# UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

MESTRADO EM: GESTÃO DE SISTEMAS DE INFORMAÇÃO

**A QUALIDADE DOS DADOS NO APOIO À TOMADA DE DECISÃO EM AMBIENTES COMPLEXOS - DATA WAREHOUSING E BUSINESS INTELLIGENCE**

ANTÓNIO CORREIA FERNANDES

**Orientação:** Mestre Eng. Aristides de Sousa Mendes

**Júri:**

**Presidente:** Doutor António Maria Palma dos Reis

**Vogais:** Doutor Fernando Manuel Pereira da Costa Brito e Abreu
Mestre Eng. Aristides de Sousa Mendes

Julho/2005

**Glossário de abreviaturas**

**ABD** - *Administrador de Base de Dados*
**AD** - *Administrador de dados*
**BD** - *Base de Dados*
**BI** - *Business Intelligence*
**CASE** - *Computer-Aided System Engineering*
**CRM** - *Customer Relationship Management*
**DECO** - *Associação de Defesa do Consumidor*
**DSS** - *Sistema de Suporte à Decisão*
**DW** - *Data Warehouse*
**EAI** - *Enterprise Application Integration*
**ED** - *Equipas de Desenvolvimento de Software*
**ERP** - *Enterprise Resource Planning*
**ETL** - *Extracção, Transformação e Carregamento de dados*
**HOLAP** - *Híbrido OLAP*
**HTML** - *Hypertext Markup Language*
**KPI** - *Key Performance Indicator*
**MOLAP** - *Multidimensional OLAP*
**OECQ** - *Organização Europeia de Controlo da Qualidade*
**OLAP** - *Online Analytical Processing*
**OLTP** - *Online Transaction Processing*
**QD** - *Qualidade dos dados*
**RM** - *Repositório de Metadados*
**ROI** - *Return On Investment*
**ROLAP** - *Relacional OLAP*
**SCIP** - *Society of Competitive Intelligence Professionals*
**SGBD** - *Sistema de Gestão de Base de Dados*
**SI** - *Sistemas de Informação*
**SQL** - *Structured Query Language*
**TDQM** - *Total Data Quality Management*
**TI** - *Tecnologias de Informação*
**TQdM** - *Total Quality data Management*
**XML** - *Extensible Markup Language*

**A QUALIDADE DOS DADOS NO APOIO À TOMADA DE DECISÃO EM AMBIENTES COMPLEXOS - DATA WAREHOUSING E BUSINESS INTELLIGENCE**

António Correia Fernandes
*Mestrado em:* Gestão de Sistemas de Informação
*Orientador:* Mestre Eng. Aristides de Sousa Mendes
*Provas concluídas em:*

## RESUMO

O gestor procura estar informado, bem informado. A informação de qualidade permite-lhe, a cada momento de decisão, reduzir as incertezas e decidir com excelência. Decidir é escolher por uma opção entre várias. Decidir com qualidade implica escolher um caminho onde, por um lado, se está consciente do que se perde no momento e, por outro, no que se pode ganhar. Vários factores intervêm no processo da tomada de decisão, sendo um deles a informação, que pressupõe dados com qualidade.

Em qualquer organização, grandes quantidades de dados relativos às diversas áreas do negócio são gerados e armazenados diariamente, passando a fazer parte do património dessa organização. Não obstante, esses dados encontram-se geralmente dispersos por várias bases de dados operacionais e, a respectiva concentração e eventual agregação enfrenta muitas dificuldades. Por um lado é necessário desenvolver sistemas adequados ao processo de centralização e, por outro, é necessário disponibilizar ao utilizador as ferramentas adequadas que lhe permitam obter resposta às suas questões.

As Tecnologias de Informação podem desempenhar um papel muito importante em todo este processo, quer na recolha, migração, armazenamento ou disponibilização de dados, quer no processo de obtenção de informação, no processo de tomada de decisão ou na partilha do conhecimento. É então fundamental todo o empenho na correcta gestão destas tecnologias, especialmente em ambientes de decisão complexos.

**Palavras-chave**: Qualidade dos Dados, Metadados, Data Warehousing, Business Intelligence, Sistemas de Suporte à Decisão e Técnicas Computacionais.

**DATA QUALITY IN DECISION MAKING SUPPORT IN COMPLEX ENVIRONMENTS - DATA WAREHOUSING AND BUSINESS INTELLIGENCE**

António Correia Fernandes
*Master's Thesis on:* Information Systems Management
*Oriented by:* Mestre Eng. Aristides de Sousa Mendes
*Concluded at:*

## ABSTRACT

The decision maker seeks to be informed, well informed. Quality information allows him to reduce uncertainty and to decide with excellence. Deciding is to choose one option among others. Quality decisions imply choosing one way where you are aware of what you loose at that moment, and what you can get. Several factors interact in this decision process. One of them is information, which results from data interpretation. At that moment, the decision maker needs quality information.

Huge amounts of data related to the different business areas are daily generated and stored, adding to organizational patrimony. Systems to centralize data are necessary. Besides, information must be made available to the users in a versatile way.

Information Technologies can play an important role in the whole process, especially in data collection, migration, storage and making it available for decision making or knowledge sharing. So, information technologies need to be correctly managed by skilled staff, especially in the complex environments.

**Key Words**: Data Quality, Metadata, Data Warehousing, Business Intelligence, Decision Support Systems and Computational Techniques.

**Índice**

## Lista de Tabelas



## Lista de Figuras

## Agradecimentos

Ao meu orientador, Mestre Eng. Aristides Sousa Mendes, pela forma como desde o início se empenhou neste trabalho, pelo apoio e disponibilidade que sempre demonstrou e pelo muito que me deu a aprender.

Aos meus amigos e colegas de mestrado Carlos Pereira, Nuno Fernandes, Pedro Moreno, Rui Monteiro e Rui Raposo.

Aos meus amigos e colegas de trabalho, especialmente à Anabela Mariño, ao Jorge Fernandes e ao Nuno Franco.

Aos meus pais, António Fernandes e Angelina Fernandes, aos meus avós e restante família o apoio e carinho que me deram ao longo de toda a vida e muito especialmente nestes últimos dois anos.

Por último, um agradecimento especial à Inês, minha esposa, pelo incentivo e apoio nas horas mais difíceis e pela valiosa ajuda na revisão desta dissertação.

# 1. Introdução

Nas organizações, os gestores confrontam-se diariamente com a necessidade de decidir e, no momento da decisão, proferem com frequência afirmações como: "...não temos dados...", "...temos tantos dados que nem sabemos como usá-los...", "... nem sabia que tínhamos esses dados ...", "...não posso esperar mais, tenho que decidir já!".

No momento de decidir a pressão é enorme e torna-se necessário decidir rapidamente, podendo levar os decisores quase ao desespero. Se os dados não apresentarem qualidade, maior será a dificuldade do gestor em "obter" a informação desejada para o auxílio à tomada de decisão.

Actualmente vive-se num mundo em que a informação é considerada um bem precioso, sendo esta um activo de grande valor no momento de tomar uma decisão. A informação obtida advém da interpretação dos dados disponibilizados ao decisor. Para se conseguir obter informação de qualidade os dados devem ser *correctos, oportunos, adequados ao negócio, relevantes* e *consistentes*, para enumerar apenas algumas das características fundamentais dos dados na obtenção de informação de qualidade. Torna-se igualmente necessário ter em conta o *conhecimento do decisor*, a sua *experiência*, as suas *crenças* e *valores*, o *meio onde está inserido* e a *satisfação pessoal*.

Contudo, as decisões tornaram-se mais complexas devido ao ambiente interno e externo das organizações, assim como ao ambiente social, económico, cultural e político. Todos estes ambientes são cada vez mais complexos e estão em constante mutação. A mudança é uma realidade que pode levar as organizações a criarem o seu departamento

especializado em planeamento, onde o principal papel passa por detectar, de forma sistemática, as exigências de mudança estrutural e desenhar novas formas de responder a tais mudanças, de modo eficaz e eficiente. O planeamento deve identificar as necessidades actuais e futuras da organização no que diz respeito a dados, a informação e ao conhecimento por forma a reflectir um alinhamento entre objectivos de negócio e as estratégias e funções dos sistemas de informação. Deve igualmente identificar os sistemas de informação com maior impacto estratégico no negócio e determinar políticas para a gestão dos recursos de informação. Deve ainda assegurar a criação de uma arquitectura de sistemas de informação que garanta a qualidade dos dados, da informação e da decisão. Para tal, tem que fornecer ferramentas de gestão específicas que ajudem os decisores a tomar as decisões.

A globalização, que transforma os mercados num mercado global, e a desregulamentação, que acaba com antigos monopólios e cria novos mercados, são algumas das tendências apontadas para as mudanças verificadas. A abolição de fronteiras e a proliferação da internet são outros dos factores que implicam mudanças nas organizações pois facilitam o aparecimento de novos concorrentes com custos de produção mais baixos e maior oferta, ou seja concorrentes fortes.

Estas transformações têm ocorrido devido à progressiva transição da sociedade pós-industrial para a sociedade do conhecimento. Na primeira, as organizações estavam habituadas a operar em meios mais estáveis, mais pequenos e protegidas pelas leis do país, tinham pouca concorrência, o que as levava a estarem praticamente centradas na gestão dos seus processos não dando atenção às mudanças que ocorriam no ambiente

externo. Actualmente, na sociedade do conhecimento, o progresso tecnológico, a explosão dos meios de comunicação e de informação, a importância das grandes multinacionais e do seu impacto nas economias bem como a globalização podem constituir uma ameaça, mas também podem ser uma oportunidade.

♦

A elaboração desta dissertação tem como base as problemáticas acima mencionadas, e neste sentido deve responder à questão de partida: *como melhorar os dados nos sistemas de apoio à decisão em ambientes complexos?* No intuito de responder a esta questão, o presente trabalho propõe um conjunto de dimensões de qualidade dos dados nos sistemas de suporte à decisão e uma arquitectura para um *Ambiente Analítico* de apoio à tomada de decisão, encontrando-se estruturado da seguinte forma:

**Capítulo dois - A Tomada de Decisão Apoiada nas TIs**: neste capítulo apresenta-se o processo de tomada de decisão e o que se entende serem os seguintes conceitos: dados, informação, conhecimento e gestão do conhecimento.

**Capitulo três - Business Intelligence**: apresenta o ambiente de Business Intelligence e suas componentes como as ferramentas de Extracção, Transformação e Carregamento de dados, o Repositório de Metadados, o Data Warehouse e as ferramentas de Business Intelligence.

**Capítulo quatro - A Qualidade dos Dados - Dados no Data Warehouse e Dados disponibilizados pelas ferramentas de BI**: tem início com uma introdução ao tema da qualidade, prossegue com uma revisão da literatura no que respeita à qualidade dos

dados finalizando, por um lado, com a proposta de dimensões de qualidade dos dados no Data Warehouse, e por outro, com a proposta de dimensões de qualidade dos dados disponibilizados pelas ferramentas de Business Intelligence.

**Capítulo cinco - Arquitectura para um Ambiente Analítico**: inicia-se com a apresentação da framework para um ambiente analítico e apresenta-se a proposta de arquitectura com algumas considerações a ter em conta na sua implementação.

**Capítulo seis - Aplicação da Arquitectura para um Ambiente Analítico numa organização concreta**: implementação e análise da arquitectura proposta num sub-sistema dos Correios de Portugal.

**Capítulo sete - Conclusão**: finaliza-se com as considerações finais relativas ao estudo realizado onde se conclui que melhorando a qualidade dos dados, com a ajuda das dimensões de qualidade dos dados propostas e implementando uma arquitectura que sustente um ambiente analítico de qualidade, torna-se mais fácil a vida dos decisores em ambientes complexos.

# 2. A Tomada de Decisão Apoiada pelas TIs

## 2.1. O Processo de Tomada de Decisão

No processo de tomada de decisão, os decisores deparam-se com um elevado número de alternativas tendo consciência que ao optar pela alternativa errada podem comprometer seriamente o negócio. Conforme refere Turban (1995), uma má decisão pode desencadear um conjunto de reacções negativas dentro da organização.

### 2.1.1. A tomada de decisão

De acordo com os estudos levados a cabo por Simon (1960), a tomada de decisão compreende três fases, sendo a primeira a fase da pesquisa de situações que necessitem de uma decisão, a segunda a fase da análise, desenho e desenvolvimento das possíveis soluções e a terceira a escolha da solução. Este autor classificou as decisões em duas categorias, decisões programadas (ambientes simples), com processos repetitivos e procedimentos conhecidos e automatizáveis, e decisões não programadas (ambientes complexos), com processos de decisão não repetitivos, regras desconhecidas e um elevado grau de incerteza.

As decisões são especialmente difíceis de tomar em ambientes complexos, uma vez que estes ambientes são caracterizados por grandes níveis de incerteza, variedade de opções e existência de pressões a vários níveis (Sousa-Mendes, 2001a). De acordo com Daellenbach (1995), esta complexidade está relacionada com a evolução tecnológica, com a proliferação dos meios de comunicação e informação e, de certa forma, com o impacto que estes podem ter em termos globais.

De facto, o actual fenómeno da globalização permite que as decisões possam repercutir-se a nível global. Segundo Guy Kolb, director executivo da Society of Competitive Intelligence Professionals (SCIP), "*O espaço para se falhar é cada vez menor, pois a pressão dos mercados e accionistas, assim como as questões de carácter social e ambiental, colocam as organizações e os seus decisores numa posição pouco propícia a falhas*" (Taborda e Ferreira, 2002, p.20). Neste contexto, a tomada de decisão torna-se um processo complexo onde o decisor necessita de instrumentos de apoio no momento de optar por uma alternativa com vista a atingir determinado objectivo.

As Tecnologias de Informação

As organizações ficam numa posição delicada no que se refere a falhas na tomada de decisão e torna-se necessário ajudar os decisores no que se refere à informação relevante (tarefa muitas vezes dificultada pelo excesso de dados contraditórios). As Tecnologias de Informação (TI) podem ser o instrumento que os decisores necessitam, uma vez que exercem um papel fundamental a vários níveis – na gestão dos dados, no processo de obtenção de informação, no processo de tomada de decisão e na contribuição para o conhecimento e respectiva partilha. Através destas capacidades, as TI, ajudam o decisor a obter informação com qualidade para o apoio na tomada de decisão.

### 2.1.2. O Sistema de Suporte à Decisão

É neste encadeamento que se insere o Sistema de Suporte à Decisão (DSS). Consiste num Sistema de Informação (SI) baseado em computadores, com grande envolvimento dos utilizadores, também eles parte do SI. Um SI, de acordo com Laudon e Laudon (1998), pode ser definido como um conjunto de componentes agregados que interagem

de forma a cumprir um objectivo definido. Retomando o DSS, este é um sistema computacional e interactivo, que apoia o decisor na tomada de decisões e recorre ao uso de modelos e dados com o objectivo de ajudar o decisor a obter informação para solucionar determinados problemas (Turban, 1995).

Características do DSS

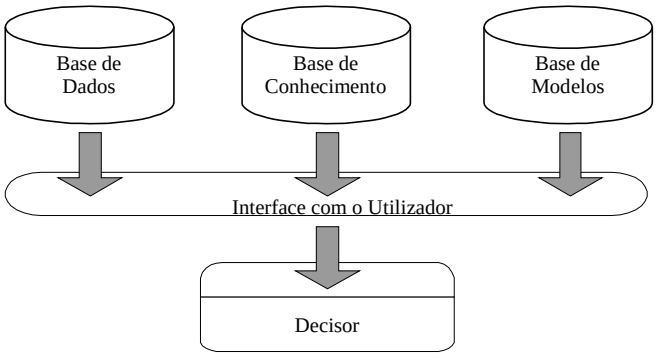
O DSS é caracterizado por ser usado no apoio às decisões mais complexas (menos estruturadas), por tentar combinar modelos ou técnicas analíticas com as funções tradicionais do processamento de dados, ser interactivo, fácil de usar e possuir interfaces amigáveis, incorporar meios mais eficientes de distribuição dos resultados (Turban, 1995; Alter, 1992). De acordo com Palma-dos-Reis (1999), um DSS deve apresentar abertura em termos de novos requisitos e novos modelos, por forma a dar resposta a novas questões colocadas pelos utilizadores. O negócio não pára, a competição é cada vez maior e a inovação tem de ser uma realidade, daí a necessidade do DSS estar aberto a novas interrogações. Neste contexto, o DSS, deve ser acessível e facilmente adaptado a novas realidades do negócio, quer pela actualização de novos dados, quer pela construção de novos modelos. Um DSS não pode ser construído e não mais sofrer alterações. Não obstante ser necessário um objectivo para desenvolver e finalizar um projecto de DSS, deve ficar clara a possibilidade de evolução desse mesmo projecto de acordo com novos requisitos de negócio que se venham a verificar.

Sub-sistemas do DSS

Um Sistema de Suporte à Decisão (DSS) é, de acordo com Turban (1995), composto por cinco sub-sistemas, são eles: o Sub-sistema de Gestão de Dados, o Sub-sistema de

Gestão do Modelos, o Sub-sistema de Gestão de Conhecimento, o Sub-sistema de Gestão de Interfaces e o Sub-sistema de Gestão de Utilizadores/Decisores. A figura abaixo espelha a forma como estes sub-sistemas estão interligados.

**Figura 1 - Sistema de Suporte à Decisão (DSS)**




Descrição dos Sub-sistemas do DSS

O Sub-sistema de Gestão de Dados inclui a Base de Dados (BD), que deverá conter os dados relevantes para a situação/problema em questão. A BD é gerida por software específico, designado por Sistema de Gestão de Base de Dados (SGBD).

O Sub-sistema de Gestão de Modelos poderá tratar modelos diversos como os estatísticos, os financeiros, de optimização, de gestão e outros modelos quantitativos, proporcionando ao sistema a capacidade analítica fundamental e o software adequado de gestão. Este software designa-se por Sistema de Gestão de Base de Modelos e, conforme refere Palma-dos-Reis (1999), uma vez construído o modelo há que assegurar a sua validade interna e externa assim como a sua manutenção. A validade interna consiste em assegurar que o modelo não apresenta erros de qualquer espécie, sejam eles erros de estrutura, erros de semântica ou erros de sintaxe. A validade externa pode ser verificada através da análise dos seus ***inputs***, dos seus ***outputs*** e dos seus pressupostos.

O Sub-sistema de Gestão de Conhecimento poderá ou não existir num DSS. Existindo, poderá ajudar qualquer um dos outros componentes ou então ser mais um instrumento para aumentar a capacidade/inteligência do DSS.

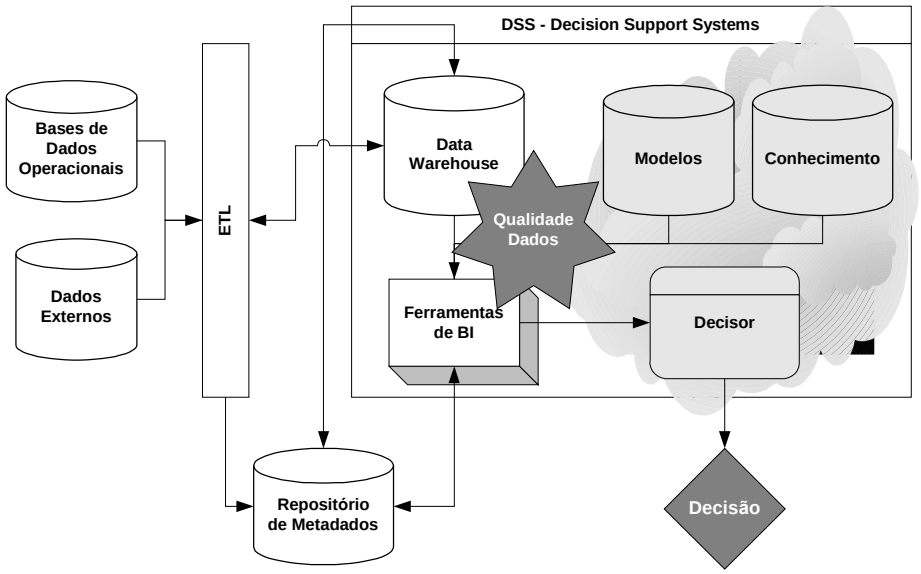
O Sub-sistema Interface com o Utilizador permite que este último interaja com o DSS. Palma-dos-Reis (1999) refere que a complexidade do DSS requer um desenho cuidadoso do interface com o utilizador, sendo necessário ter em atenção alguns cuidados no sentido de facilitar a comunicação entre o decisor e o resto do sistema.

O Sub-sistema Utilizador/Decisor é composto pelo próprio decisor, que é de facto um Sub-Sistema do DSS dada a sua forte e imprescindível interacção. O DSS não substitui o decisor na tomada de decisão, logo o decisor faz parte do sistema e tem como função interpretar os dados disponibilizados pelo DSS.

Arquitectura do DSS

Como se refere na introdução, esta dissertação tem como primeiro objectivo o melhoramento da qualidade dos dados armazenados no Data Warehouse (DW) e disponibilizados pelas ferramentas de BI, que correspondem, respectivamente, à componente Modelo de Dados e à componente Interface com o Utilizador. Na figura 2, ilustram-se os outros componentes do DSS assinalados com uma nuvem. Pretende-se salientar a sua importância, contudo, no âmbito deste trabalho apenas é feita referência a estes componentes para fins de contextualização.

**Figura 2 - Arquitectura DSS**




## 2.2. Dados, Informação e Conhecimento

Ao desenvolver a temática do processo de tomada de decisão, torna-se evidente a necessidade de clarificar conceitos como *Dados* e *Informaçã*o, muitas vezes referenciados como se do mesmo significado se tratasse. Será igualmente objecto de reflexão o entendimento dos conceitos *Conhecimento* e *Gestão do Conhecimento*.

Dados e informação (pesquisa bibliográfica)

Vejamos o que a este respeito encontramos na bibliografia consultada. Huang, Kuan-Tsae *et al.* (1999, p.13), na obra *Quality Information and Knowledge* referem:

"*Os termos Dados e Informação são muitas vezes utilizados como sinónimos; na prática, os gestores diferenciam-nos intuitivamente, e entendem o termo informação como dados que tenham sido processados.*"

Os dados representam coisas ou entidades do mundo real, são a matéria em "***bruto***" da qual deriva a informação. A informação é o produto final, isto é, a contextualização e o significado dos dados para que os factos se tornem inteligíveis (English, 1999).

Davenport e Prusak (1998) consideram os dados como observações sobre o estado do mundo. São símbolos e imagens que não reduzem as nossas incertezas. São elementos em bruto, desvinculados da realidade e que constituem a matéria-prima da informação. Se não tiverem qualidade, a informação que deles resulta também não terá.

Segundo Rascão (2000), dados são factos, eventos, imagens ou sons que podem ser pertinentes ou úteis para o desempenho de uma tarefa mas que por si só não conduzem a uma compreensão de determinado facto ou situação.

Os dados só poderão ser úteis se tiverem significado e quando contextualizados, caso contrário o que se observa é apenas *ruído*. Isto é, "20" por si só não é um dado, é *ruído*. Vejamos o mesmo "20" mas num determinado contexto: o António mora na rua Principal, número 20. Continuamos com um dado, que agora, além de ter significado, está inserido num contexto: o dado diz-nos que o António mora na rua Principal, número 20. Neste caso podemos servir-nos deste dado para obter informação. Contudo, se tivermos presente o dado *código postal* contextualizado, código postal 2380 Pousados, então conseguimos obter informação mais rica, neste caso, a informação de que: o António vive na rua Principal, número 20 com o código postal 2380 Pousados. Esta informação que acabamos de alcançar é mais completa, ou seja, mais redutora de incerteza.

A representação de factos ou fenómenos é feita através de *dados*, não de informação. Mas para entender a realidade à sua volta e para tomar decisões, as pessoas necessitam de *informação* (Sousa-Mendes, 2001a). A informação é redutora de incerteza, e quanto menor for a incerteza melhor se pode entender a realidade à nossa volta e mais fácil será tomar uma decisão. Ilharco (2003) considera que informação pode ser vista como "*a diferença que faz a diferença"*.

Segundo Le Moigne, citado por Rascão (2001, p.21), informação é "*um objecto formatado pelo homem, tendo por finalidade representar um tipo de acontecimento identificável por ele no mundo real, integrando um conjunto de registos ou dados e um conjunto de relações entre eles, que determinam o seu formato*".

Para Ilharco (2003, p.36), "(...) *a forma como simples dados se transformam em informação, é que varia não apenas de pessoa para pessoa, mas também de instante para instante*". Este autor refere ainda que a "*Informação é o significado para o sujeito que experimenta a acção de ser/estar/ficar informado. Nesta perspectiva a informação é um fenómeno interpretativo, dependente do sujeito, assente na experiência de determinado indivíduo e na historicidade, pressupostos, contextos e envolvimentos no âmbito dos quais e com os quais esse mesmo indivíduo se informa ou é informado*".

### 2.2.1. Os Dados

Existem efectivamente diferenças entre dados e informação, sendo que os dados são objectivos, têm sempre significado e não reduzem a incerteza. Podem ser encontrados em várias formas (vídeo, relatórios, gráficos,...) e são a matéria-prima da informação. Sousa-Mendes (2001b) refere, a título de exemplo, que um gráfico está longe de ser

informação, e a prova está que para ser bem interpretado é necessária alguma formação. Um dado transforma-se em informação quando ganha significado para o utilizador num determinado contexto, caso contrário, continua apenas a ser um dado.

O Ciclo de Vida dos Dados

O ciclo de vida dos dados está associado a um processo permanente que, ao usar dados para obter informação, contribui de certa forma para enriquecer a base de dados, uma vez que o aproveitamento da informação gerada fornece, quase sempre, novos dados. Os dados circulam num ciclo e geram cada vez mais valor, pois em cada ciclo são melhorados ou refinados. De seguida, descreve-se o ciclo de vida dos dados desde a sua *definição*, momento em que se prepara a sua recolha, até à sua *utilização*, momento em que são utilizados para a obtenção de informação.

Segundo Huang, Kuan-Tsae *et al*. (1999), na obra *Quality Information and Knowledge*, informação é um produto do qual os dados são a matéria prima, devendo ser gerida a informação tal como se gere um produto. Pode-se, a partir daqui, fazer um paralelo entre o ciclo de vida dos dados e o ciclo de vida dos produtos. Submetendo o ciclo de vida de um dado a um raciocínio biológico, apontamos quatro fases: o nascimento, que corresponde ao lançamento do produto; o desenvolvimento, a sua ascensão para o estádio seguinte; a maturidade, que se verifica no seu auge; e por fim o declínio, que corresponde à velhice, e consequentemente conduz à morte. Considera-se que os dados podem ser sempre úteis, independentemente da idade, uma vez que o seu valor depende do contexto em que vão ser usados. Assim, defendemos que os dados constituem um

contributo imprescindível para a obtenção de informação sendo, contudo, necessária uma prévia selecção de acordo com o objectivo a atingir.

O ciclo de vida dos dados tem início no momento em que a sua necessidade é identificada. Esta fase engloba a *definição* e *estrutura dos dados* armazenados. Devem ser definidos os tipos de dados a usar, as regras de validação e de integridade, o tamanho, os nomes, o domínio dos dados. Estas questões prendem-se com metadados, alvo de estudo mais à frente neste trabalho. Se nesta fase for tida em conta a preocupação de garantir qualidade, mais tarde ganha-se com isso, pois a necessidade de processos de "*limpeza*" de dados será menor. Muitos dos factores que contribuem para a falta de qualidade dos dados já terão sido identificados e tidos em conta nesta fase. Estas preocupações serão ponderadas mais à frente no capítulo dedicado à QD.

O processo continua com a **Recolha** dos dados. A recolha pode ter origem em fontes Internas – actividades internas, escritório, e Externas – artigos, jornais, revistas, bases de dados comerciais ou mesmo dados dos concorrentes. A recolha dos dados é muito importante, uma vez que estes devem ter origem em fontes credíveis e ser recolhidos de forma adequada.

Numa fase seguinte, os dados devem ser armazenados numa base de dados, em sentido genérico (**Armazenamento**). Após identificadas as necessidades, identificados os dados necessários e feita a sua recolha ou carregamento, estes devem ser guardados para posteriormente serem acedidos, processados e disponibilizados a quem deles necessite.

O **Processamento** de dados decorre da necessidade de trabalhar os dados para que estes sejam facilmente interpretados pelo decisor. Se deste processo resultarem novos dados, estes também devem ser armazenados. Todos os dias surge a possibilidade de aumentar a quantidade de dados na base de dados, quer através de novas recolhas ou carregamentos, quer pelo seu processamento.

A fase da **Disponibilização** ocorre após o processamento dos dados. Nesta fase os dados são disponibilizados ao decisor para que este obtenha, através da sua interpretação, informação com vista à tomada de **decisão/acção**.

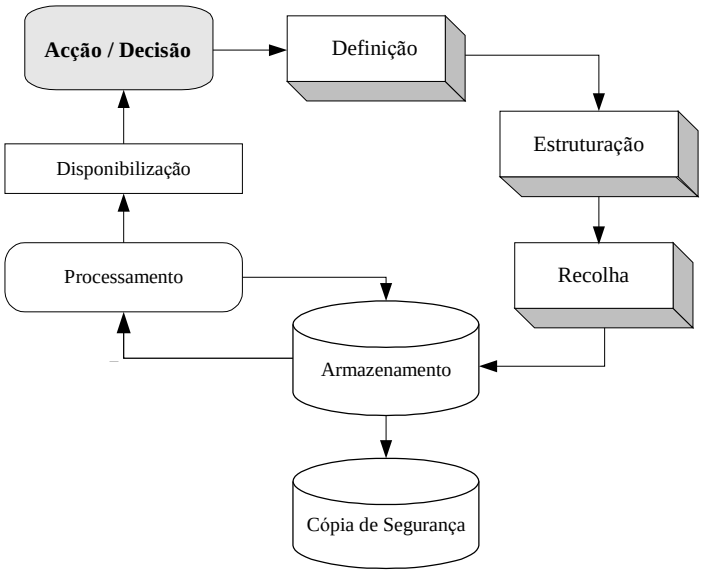
A Gestão dos dados

Reforça-se a ideia de que os dados representam um grande valor para a capacidade de obtenção de informação e de conhecimento. Sem tal capacidade a organização enfrentaria sérios problemas de sobrevivência. É de realçar a questão da importância de uma cópia de segurança dos dados (*backup*), assim como de uma política de controlo de acessos que impeça o acesso aos dados por parte de agentes não autorizados.

Uma questão a ter em conta prende-se com a decisão de manter os dados em produção ou passá-los para histórico, uma vez que a eliminação definitiva de um dado pode ser prejudicial à organização. Num determinado momento pode decidir-se eliminar alguns dados que pela sua idade, falta de uso, ou outro factor já não interessem à organização. Contudo, deve-se considerar a hipótese de, mais tarde, outro decisor ter a necessidade de alguns indicadores referentes a dados antigos, sendo que a ausência destes pode tornar a tarefa bastante difícil ou mesmo impossível. Neste caso, dados mais antigos

podem auxiliar previsões, como por exemplo o evoluir do negócio com base em situações passadas: os dados das vendas de 1960 podem ser úteis na obtenção de certos indicadores, os dados do ano 2002 podem ser úteis para a obtenção de outro tipo de indicadores, ou até mesmo reforçar os atrás descritos, conjugando ambos os indicadores para chegar ao indicador pretendido. Estes cenários devem ser considerados antes de se decidir pela eliminação de qualquer dado. A figura abaixo ilustra o ciclo de vida dos dados.

**Figura 3 - Ciclo de Vida dos Dados**




### 2.2.2. A informação

A informação, por sua vez, é o sentido atribuído aos dados, após o seu processamento e interpretação com o objectivo de ajudar na tomada de decisão. A informação é subjectiva e o seu valor está em reduzir a incerteza, sendo tanto mais rica quanto mais possibilidades forem excluídas. Contudo, a obtenção de informação também depende de conhecimentos anteriormente adquiridos. A informação validada pode ser uma das matérias-primas do conhecimento pois, como refere Sousa-Mendes (2001a), o

conhecimento é o resultado da validação e integração de informação num receptor, com o sentido de utilidade para um determinado fim.

Informação de qualidade

Segundo Ilharco (2003), "*ver é determinante para a acção"*. A forma como vemos depende, entre outras coisas, de quem somos e do que sabemos. Thomas Kuhn, citado pelo mesmo autor, defende que "*o que um homem vê depende da forma como olha para o quê e também daquilo que a sua prévia experiência visual e conceptual o ensinou a ver*". A forma como o decisor interpreta os dados depende do que ele é como pessoa, do seu conhecimento, do seu capital cultural, do seu nível intelectual, da sua motivação. Assim, a interpretação dos dados gera, por parte do decisor, informação com maior ou menor qualidade. Para a qualidade da informação em muito contribui o facto de se terem dados de qualidade e um decisor capaz de os interpretar.

No fundo, a boa informação, ou informação de qualidade, resulta fundamentalmente de quatro factores: os dados, o tratamento a que estão sujeitos (software), o receptor e o contexto em que o receptor se confronta com os dados (Sousa-Mendes, 2001b). Este autor refere ainda que a informação resulta da interpretação de dados, dividindo-os nos elementares e nos complexos, e salienta que os dados são o único elemento objectivo em todo o processo de obtenção de informação. A interpretação dos dados tem como base a sua qualidade mas também o contexto em que são interpretados, e este contexto tem que ver com um conjunto de factores dos quais se destaca o posto de trabalho, o ambiente cultural da organização, a conjuntura interna e externa e o próprio software

com que o decisor opera (quer ao nível de *front-end*, quer do próprio processamento de dados).

### 2.2.3. O Conhecimento

"*Conhecimento é uma mistura fluida de experiência enquadrada, valores, informação contextual e compreensão especializada que fornece um quadro para avaliação e incorporação de novas experiências e informação. É originada e aplicada nas mentes dos seus detentores. Nas organizações aparece muitas vezes embutida não apenas em documentos e repositórios mas também nas rotinas, processos, práticas e normas.*"

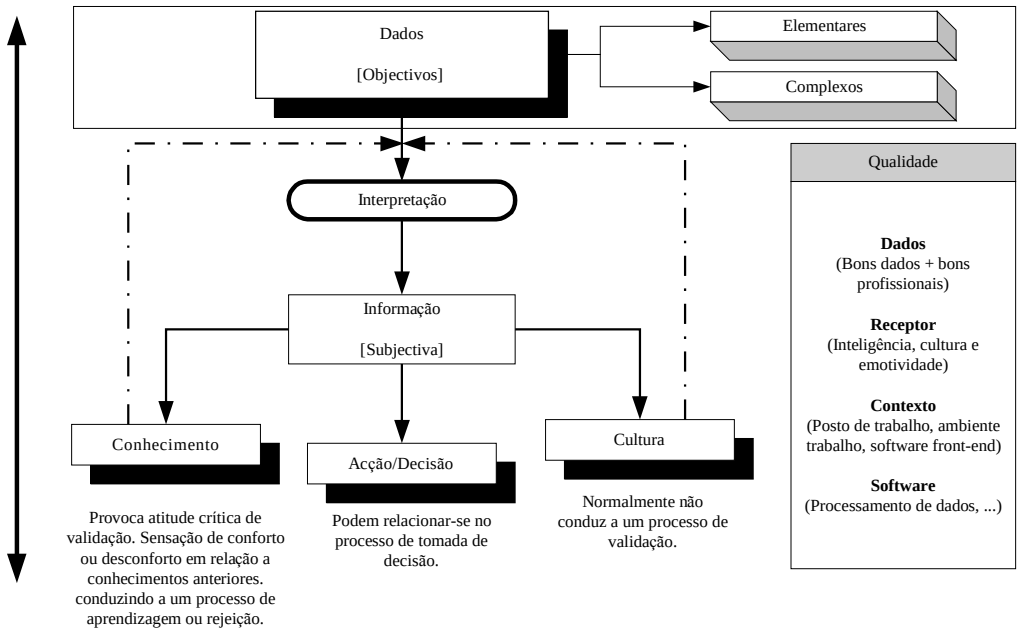
Davenport e Prusak (1998)

Conhecimento é, de acordo com Tiwana (2000), "*informação para a acção*", informação relevante. O conhecimento é informação validada e armazenada que pode ser acedida para suporte ao processo de decisão. O factor conhecimento contribui igualmente para uma melhor interpretação, uma vez que a informação obtida é tanto melhor quanto maior for o grau de conhecimento da pessoa que interpreta os dados.

Presentemente o conhecimento é considerado o recurso económico mais importante, mais ainda que os recursos naturais, a mão-de-obra ou o capital (Drucker, 1996). A capacidade intelectual pode potenciar novas ideias, novas formas de resolução de problemas assim como interpretar factos relevantes, agregando, desta forma, valor às organizações. Igualmente, o conhecimento de um indivíduo é fundamental no processo que este acciona ao transformar dados em informação relevante. No entanto, não se deve esquecer a dose de subjectividade inerente à informação.

O conhecimento adquire-se pela validação da informação obtida, que pode servir para tomar uma decisão mas também, como salienta Sousa-Mendes (2001b), para aumentar o grau de conhecimento da pessoa que interpreta os dados, desde que esta informação tenha a ver com a sua área de actuação. Caso não se relacione directamente com a área de actividade do decisor, a informação obtida contribui para aumentar o seu nível cultural e, neste caso, é mais difícil validar. O esquema representado na figura 4 ajuda a perceber como interagem os dados, a informação, o conhecimento e os decisores no processo de tomada de decisão.

**Figura 4 - Dados, Informação e Conhecimento**




O conhecimento, conforme referido, é obtido da validação e agregação da informação (Sousa-Mendes, 2001a). Para ambas as operações é necessária a participação humana. Neste sentido, as pessoas e a sua experiência, além dos investimentos em tecnologia, têm muito valor para a organização. A tecnologia, entre outras coisas, serve como meio facilitador para que se possam obter, partilhar e usar dados e informação. Contudo, o

factor humano é muito importante e insubstituível no processo de obtenção de informação ou conhecimento.

De acordo com Davenport e Prusak (1998), as bases do conhecimento são: *Experiência, Juízo, Regras básicas, Intuição e Valores.* A forma clássica de tratar o conhecimento consiste na distinção entre conhecimento *tácito* e conhecimento *explícito* (Nonaka e Takeuchi, 1995).

*Conhecimento tácito* – competências, juízos e intuições que as pessoas possuem mas que não são facilmente descritos. Pode traduzir-se pela frase: "A pessoa sabe mais do que aquilo que é capaz de transmitir". O conhecimento tácito é subjectivo, prático e automático, uma vez que exige pouco ou nenhum tempo de reflexão para o aplicar. Porém, pode apresentar problemas como: estar errado; ser difícil de modificar; ser difícil de comunicar.

*Conhecimento explícito* – competências e factos susceptíveis de serem documentados e formalmente transmitidos. Pode ser facilmente expresso por palavras ou números e pode ser pronta, formal e sistematicamente transmitido entre pessoas. Envolve o conhecimento de factos e é objectivo.

É comum afirmar-se que o conhecimento reside nas pessoas e é mais rico que a informação por si só. Não se pode descurar o facto de o conhecimento ser um trunfo muito forte, encarado até como uma grande vantagem competitiva, mas tudo depende da perspectiva de gestão a que está sujeito por parte da organização.

### 2.2.4. A Gestão do Conhecimento

Quinn (1992) afirma que o poder económico e de produção de uma empresa moderna está mais centrado nas suas capacidades intelectuais e de prestação de serviços do que nos seus activos imobilizados, como a terra, as instalações ou os equipamentos. Prevê-se que as organizações do futuro terão no conhecimento, e na sua gestão, uma das grandes vantagens competitivas. As que melhor saibam gerir está área, mais aptas estão para competir. O conhecimento pode ser obtido através de indivíduos ou grupos detentores de conhecimento e a organização deve ter uma política capaz de reter tais conhecimentos dentro dela. Como tal, terão de dar especial importância à gestão do conhecimento e à maneira de gerir os seus recursos humanos para que estes se sintam motivados e partilhem os seus conhecimentos. As organizações devem refinar todo o conhecimento existente e armazená-lo para que o possam utilizar mais tarde.

Anthony (1965) defendia a teoria de que a empresa poderia ser vista como uma pirâmide, gerida por departamentos estanques. Mais tarde, adoptou-se a gestão por projectos (exemplo das linhas de montagem de automóveis). Nos anos 90 surgiu a gestão por processos, resultado de um conjunto de técnicas que ficaram conhecidas como "reengenharia de processos". Os resultados nem sempre foram os melhores, existindo casos de insucesso (Alves, 1995). Uma implicação muito forte da reengenharia de processos é a mudança organizacional, e esta nem sempre é bem explicada aos colaboradores, comprometendo assim o sucesso da reengenharia. É necessário uma gestão da mudança muito bem planeada para chegar a "***bom porto***".

"*A gestão eficaz do conhecimento só poderá ocorrer com a ampla mudança comportamental, cultural e organizacional. A tecnologia isoladamente não fará com que a pessoa possuidora do conhecimento o compartilhe. A tecnologia isoladamente não levará o funcionário a sentar-se diante do teclado e começar a pesquisar. A mera presença de tecnologia não criará uma organização de aprendizagem contínua nem uma empresa criadora de conhecimento*."

(Davenport e Prusak 1998, citado por Baroni *et al.*, 2003, p.215)

Numa organização direccionada para a aquisição e partilha do conhecimento, as funções passam a ter uma componente intelectual no desenvolvimento do trabalho desempenhado por todos os colaboradores. Como é referido por Sousa-Mendes (2001a), todos os colaboradores têm uma componente operacional e outra de gestão no desenvolvimento das suas actividades. Tal implica que os trabalhadores têm de saber operar com equipamento sofisticado e ter capacidade de tomar decisões. Como consequência, dá-se um achatamento da hierarquia.

Neste cenário é importante reconhecer o *Capital Humano* como um dos principais meios geradores de riqueza das organizações, e este reconhecimento é já uma das fontes de mudança proporcionadas pela Sociedade do Conhecimento. A boa gestão do *Capital* pode potenciar activos de grande valor para a organização, os *trabalhadores do conhecimento.* Junqueiro (2002) salienta o facto dos referidos trabalhadores, através do seu conhecimento, transformarem informação técnica e especializada em acção.

A exigência sobre estes trabalhadores está a aumentar. Vive-se numa sociedade cuja única certeza é a mudança e onde tudo muda a toda a hora é impossível estar-se munido de uma receita de sucesso. Actualmente, um profissional está rodeado de paradoxos como: pensar a longo prazo mas obter resultados imediatos; inovar sem perder a eficiência; colaborar mas também competir; trabalhar em equipa mas ser responsabilizado individualmente; conviver com um real cada vez mais virtual; estar focado num ponto sem perder a noção do que o rodeia; ser rápido e agressivo mas com consciência e emoção (Rezende, 2002).

Estas mudanças têm como resultado a necessidade de profundas alterações na gestão da organização quanto à forma, ao processo de gestão e ao estilo de liderança. A organização deve facilitar a formação de grupos de aprendizagem, concentrando-se em questões de desenvolvimento dos colaboradores (Senge, 1990).

A aprendizagem organizacional

Promover a aprendizagem organizacional é fundamental. O conhecimento individual não chega, é necessário que haja partilha do conhecimento e, neste caso, a cultura organizacional é um factor crítico de sucesso, pois sem ela é difícil atingir tal objectivo. Para que as organizações recolham os frutos, proporcionados pela Sociedade do Conhecimento, é necessário que os gestores disponibilizem boas condições aos seus colaboradores, proporcionem o bem-estar e a satisfação pessoal e acima de tudo, criem um ambiente de confiança onde todos se sintam parte integrante do projecto. Os trabalhadores devem sentir que fazem parte da estratégia de forma a contribuírem para o alcance dos objectivos da organização, só assim se pode ter sucesso na partilha do

conhecimento e evitar que os trabalhadores vejam na manutenção e retenção deste uma forma de garantir o posto de trabalho. É necessário uma comunicação clara e atempada para que a organização caminhe no sentido da nova Cultura Organizacional em que a partilha do conhecimento seja um factor chave para o sucesso.

"*O papel dos gestores na Nova Economia é a criação de um clima que permita aos trabalhadores do conhecimento a aprendizagem - a partir da sua própria experiência, uns com os outros, com os seus clientes, fornecedores e parceiros de negócio*".

(Webber 1993, citado por Silva *et al.*, 2003, p.202)

A partilha do conhecimento

A partilha do conhecimento é hoje uma realidade, basta verificar os sítios na internet especializados em diversos temas como por exemplo a saúde. Um médico pode ter uma dúvida sobre como interpretar um valor observado nas análises clínicas dum paciente e pedir opinião aos seus colegas usando um fórum de discussão. Ao obter resposta, ele pode fazer uma interpretação dos resultados muito mais fiável, pois tal interpretação é feita com base nos seus conhecimentos e nos conhecimentos dos seus colegas.

Esta é a sociedade para onde caminhamos e que pode ser muito produtiva. Tudo depende da capacidade de mudarmos a forma de estar e pensar, e isso consegue-se se formos capazes de induzir, em nós próprios e naqueles que nos rodeiam, um conjunto de valores, entre eles os valores éticos. A partilha do conhecimento em muito pode contribuir na melhoria da qualidade de vida de cada um de nós. Contudo, ainda é necessário percorrer bastante caminho, principalmente na forma como as organizações terão de conciliar o conhecimento dos seus colaboradores, a informação que estes obtêm

das bases de dados, dos papéis, dos relatórios e dos gráficos, de forma a converter tudo isto em vantagens estratégicas para o negócio e, ao mesmo tempo, reter esse enorme potencial dentro da organização.

Actualmente, a mão-de-obra encaminha-se de forma gradual para a *inteligência-de-obra*[1] uma vez que o esforço intelectual tende a superar o esforço físico no desempenho das tarefas, conduzindo, desta forma, a profundas alterações no modo de gerir. Esta realidade é, em parte, fruto do uso das tecnologias de informação e comunicação e cabe à sociedade contemporânea colocar essas tecnologias ao seu serviço, tendo como objectivo maior o aumento da qualidade de vida.

[1] Este termo baseia-se no conceito *cérebro-de-obra*, utilizado por Angeloni (2002, p.21).

# 3. Business Intelligence

## 3.1. O ambiente de Business Intelligence

Business Intelligence (BI), termo idealizado pelo Gartner Group no relatório do mês de Setembro de 1996, pode ser encarado como um ambiente ou cultura decisional onde por um lado, se incentiva o uso de uma arquitectura capaz de comportar grande quantidade de dados com qualidade, e por outro se disponibilizam esses dados ao gestor garantindo a mesma qualidade (SAS Institute, 2004). O BI é um ambiente detentor de capacidade para entender o passado, compreender o presente e antecipar o futuro. Este ambiente permite analisar os dados de forma a compreender as consequências de determinada decisão ou perceber como estão a decorrer as actuais, ou ainda antever o impacto que terão as futuras. De acordo com White (2003), o ambiente de BI permite transformar dados brutos em dados relevantes, precisos e úteis, ajudando ainda o decisor a convertê-los em informação de qualidade para a análise ou tomada de decisão. Ainda segundo este autor, o ambiente de BI permite analisar ou tomar decisões com maior rapidez e segurança.

O Ambiente BI

Este ambiente envolve várias tecnologias, nomeadamente, as ferramentas de ETL, o Data Warehouse (DW), o Repositório de Metadados (RM) e as próprias ferramentas de BI (Rubin, 2003). Envolve, ainda as Pessoas, os Processos e a Cultura organizacional, permitindo, desta forma, uma vigilância permanente sobre o negócio para que, quando necessário, os decisores possam intervir tendo presentes os dados correctos e em tempo.

### 3.1.1. As ferramentas de BI

Estas ferramentas têm como objectivo a disponibilização de dados relativos ao negócio. Podem traduzir-se na capacidade de ter um controlo global sobre o negócio e, desta forma, levar ao aumento da credibilidade da organização ou mesmo, à obtenção de vantagens competitivas. De acordo com a ORACLE (2004), a utilização de ferramentas de BI deve estar alinhada com os objectivos estratégicos da organização. As ferramentas de BI facilitam a consolidação de dados presente na organização, tornando, desta forma, mais fácil o planeamento e a execução de tarefas. A SAS Institute (2004) refere ainda que, com o uso de ferramentas de BI, as tarefas da organização passam a ser realizadas com superior agilidade, confiança, monitorização e transparência.

Funcionamento das ferramentas de BI

As ferramentas de BI trazem para as organizações características da inteligência humana, com o intuito de reproduzirem padrões humanos de comportamentos e atitudes inteligentes quando confrontadas com problemas a resolver. As organizações para terem um melhor êxito nas suas decisões não podem, de forma alguma, descurar a mais valia proporcionada por tais ferramentas (SAS Institute, 2004). Para tal, não podem desprezar os dados que têm em seu poder, os que não têm de momento, mas que estão ao seu alcance, e socorrendo-se das tecnologias de informação apropriadas, gerir esses dados, de forma a ser possível o seu uso pelos decisores no momento de obter informação.

Uma ferramenta de BI disponibiliza dados de várias naturezas como, de mercado, de produtos, de concorrentes, de clientes, de fornecedores, de processos, de tecnologias. Alguns autores referem o papel crítico desempenhado por estas ferramentas no actual

contexto competitivo, quer pela exigência cada vez maior dos consumidores, quer pela globalização dos mercados (Davenport e Prusak, 1998; Nonaka e Takeuchi, 1997).

As ferramentas de BI, de uma forma geral, disponibilizam os dados através de Browser e, permitem exportá-los em determinados formatos, como: gráficos, documentos pdf, documentos Word e Excel. Facilitam a gestão e acessos a conteúdos, recorrendo ao perfil de utilizador (ORACLE, 2004).

Com o uso destas ferramentas, a organização, passa a dispor de uma panóplia de soluções, de alto nível tecnológico, que a apoiará sobremaneira no processo de tomada de decisão. Desde a década de 90, têm surgido no mercado várias tecnologias e novas ferramentas, com o objectivo de evoluir e aperfeiçoar as ferramentas de BI já existentes.

### 3.1.2. Data Warehouse

Para Devlin (1997) um DW é um armazém consistente e completo de dados obtidos de diferentes origens. Poe *et al*. (1998), consideram que um DW é uma base de dados analítica usada pelo DSS e desenvolvida para grandes volumes de dados, armazenados de forma a melhor responder a análises multidimensionais (existência de redundância de dados). Complementando com outra ideia, no DW devem estar replicados todos os dados, extraídos das BD operacionais, necessários aos processos de tomada de decisão (Inmon, 1997; Kimball, 1998). O DW pode então ser visto como um armazém de dados integrado e único que disponibiliza a base estrutural para as aplicações de apoio à decisão usadas na organização.

Características

Segundo Inmon (1997), um DW deve ser visto como uma colecção de dados *orientada para os assuntos do negócio da empresa*, *integrada*, ou seja, construída por integração de fontes de dados múltiplas e heterogéneas. Deve ser *variável com o tempo*, já que o horizonte de tempo para um DW (5 a 10 anos) é significativamente maior do que nas BD operacionais (dados recentes até 60 a 90 dias). Nas BD operacionais os dados podem ser modificados, já no DW depois de carregados, os dados não podem ser modificados, encontrando-se armazenados de forma histórica e apenas fazem sentido se contextualizados com a data da sua criação. O DW deve ainda ser *não volátil*, ou seja, um repositório fisicamente separado das aplicações operacionais e sem actualizações directas de dados por parte destas. Requer apenas duas operações nos dados, o carregamento e o acesso. Com estas condições o DW permite um acesso fácil e rápido aos dados organizacionais. Estes são consistentes, podem ser combinados e reconvertidos para analisar variáveis de negócio.

A construção de um DW

A Construção de um DW é um processo iterativo e contínuo, e não um projecto fechado, devendo ser uma plataforma de desenvolvimento modular (incremental) e escalável (Kimball e Ross, 2002). A participação na construção do sistema é um factor crítico de sucesso sendo necessário que a organização se envolva e se faça representar ao mais alto nível. Conforme refere Kimball (1998), para o sucesso de um projecto de DW é necessário grande envolvimento da organização e empenho da gestão. Ainda segundo este autor, o DW é de toda a organização e devem ser criadas condições para que todos lhe possam e devam aceder.

Das características dos dados armazenados no DW, destacam-se de seguida duas: a granularidade e a agregação.

Granularidade

A *Granularidade* diz respeito ao nível a que os dados do DW estão sumarizados (Kimball 1998). O "grão" é o nível mais detalhado do dado. Se o grão for definido no seu nível mais detalhado, então o utilizador poderá ver esse dado em qualquer nível da agregação. No entanto há que ter em atenção que quanto mais baixo for o nível, mais espaço de armazenamento é necessário no DW, podendo afectar a sua performance. Se o grão for definido com pouco detalhe, o utilizador ficará impossibilitado de realizar consultas com grande nível de detalhe. É necessário um equilíbrio para não comprometer o trabalho do utilizador nem a performance do DW.

Agregação

A forma de não comprometer o tempo de execução de uma consulta complexa pode ser agregar os dados através de estruturas hierárquicas. Kimball (1998) e Poe *et al.* (1998) referem que os agregados são criados para aumentar o desempenho das consultas, existindo casos em que os ganhos podem chegar a um factor de 100 a 1000 vezes superiores. Poe *et al.* (1998), referem ainda o facto de os agregados reduzirem o número total de ciclos no CPU da máquina. Não é necessário ter uma preocupação exaustiva sobre quais os dados a agregar no momento de criar o DW, isto porque, podem ser acrescentados novos agregados à medida das necessidades.

Modelos de dados

Um modelo é uma abstracção do mundo real. Um modelo de dados é então uma representação das características do mundo real, que se pretende vir a tratar informaticamente, dentro de uma área específica ou assunto. Conforme salienta Singh (1997), o modelo de dados, tem como objectivo reflectir o significado dos dados, o relacionamento entre eles, os seus atributos e as suas definições.

A modelação de dados é um método que permite estruturar os dados de acordo com as regras ou conceitos de negócio de forma a possibilitar uma melhor análise futura destes. Para Ballard e Herreman (1998), inicialmente a modelação de dados, nos sistemas de apoio à decisão, não era tão importante porque estes sistemas se adaptavam aos modelos dos sistemas operacionais. Presentemente, este é um processo muito importante, com implicações no bom ou mau desempenho de um DSS.

Kimball (1998) aponta a existência de uma grande diferença no modelo de dados das BD operacionais e do DW. De facto, os modelos de dados são diferentes e com diferentes objectivos. Enquanto nos sistemas táctico-operacionais o que conta é a velocidade com que se acede e actualiza um dado ou um conjunto de dados (eficiente), no DW, o que conta é a possibilidade e facilidade de dispor de dados agregados com um nível de agregação conveniente (eficaz).

Modelo Entidade-Relacionamento

Uma das técnicas na modelação de dados mais conhecidas é o modelo Entidade-Relacionamento proposta por Chen (Chen, 1976). Esta abordagem tornou-se universal

para a modelação de dados, devido ao facto de ser simples e poderosa.

O desenvolvimento do modelo relacional teve por base a teoria dos conjuntos. Em 1970, E. F. Codd publicou um artigo com os fundamentos teóricos do modelo relacional (Codd, 1970). Contudo, apenas em 1979/80 surgiu no mercado o primeiro produto com características relacionais (SGBD Oracle). A aceitação deste modelo não foi fácil, sendo que algumas pessoas defendiam mesmo que os sistemas relacionais nunca iriam levar a uma utilização comercial. Por outro lado, os defensores deste modelo argumentavam tratar-se de um modelo muito simples e flexível (Pereira, 1998). Com o decorrer do tempo este modelo foi-se propagando pelos mercados, e neste momento está presente em quase todas as organizações onde se utilizem as tecnologias de base de dados.

Um modelo Entidade-Relacionamento procura eliminar a redundância nos dados. O que, segundo Kimball (1998), é muito benéfico no processamento dos dados. Este modelo utiliza dois conceitos: as entidades e os relacionamentos entre essas entidades. O modelo, de forma mais detalhada, contém ainda os atributos, ou seja, as características dessas entidades e relacionamentos. Cada entidade lógica dá origem a uma tabela física na base de dados.

Modelo Multidimensional

A modelação multidimensional permite apresentar os dados de forma estruturada e intuitiva para que o seu acesso seja feito com grande performance. Este modelo, segundo Kimball (1998), contém três conceitos básicos: factos, dimensões e métricas.

Um facto é uma colecção de itens de dados relacionados. Regra geral, cada facto representa um item ou transação de negócio. Os factos são inseridos na tabela de factos e podem apresentar-se de forma simples ou agregada, podendo estes conduzir a ganhos de performance.

As tabelas dimensionais armazenam as descrições textuais das dimensões de negócio (Balard e Herreman, 1998). Cada tabela de dimensão tem uma chave primária que corresponde a um dos componentes da chave composta da tabela de factos. A métrica pode ser definida como um atributo numérico de um facto, representando o comportamento do negócio. As métricas podem ser vendas em dinheiro, volumes de vendas, quantidades fornecidas, quantidades aceites entre outras.

Para Singh (1997) este modelo, conhecido por modelo em estrela, apresenta algumas vantagens importantes:

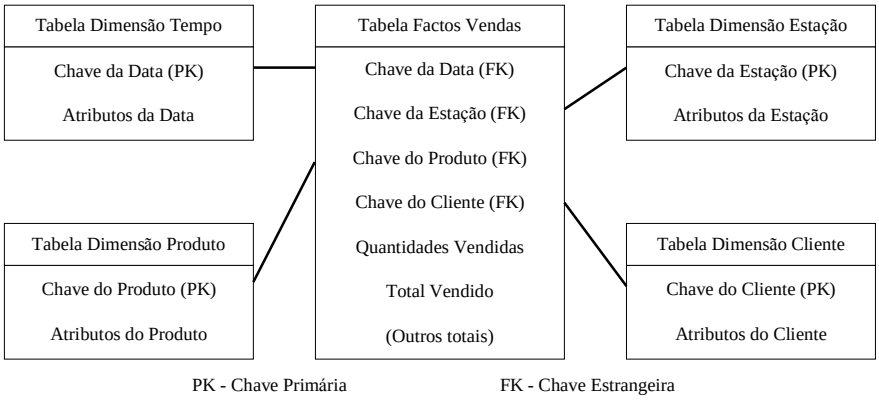
- Permite que uma estrutura de dados complexa seja representada com um modelo de dados muito simples.
- Facilita a definição de relacionamentos hierárquicos dentro de cada dimensão, simplificando a tarefa de criar as junções entre as tabelas.
- Reduz o número de junções físicas e de pesquisas, aumentando a performance.
- Devido à forma como é apresentado visualmente, facilita os utilizadores na sua interpretação, reduzindo a possibilidade de construção de pesquisas erradas.
- Permite a expansão e desenvolvimento do DW com baixa manutenção.

Exemplo do Modelo em Estrela

Para facilitar a compreensão deste modelo, aplica-se de seguida um exemplo

relacionado com o negócio dos Correios, mais propriamente a venda de objectos postais. Este modelo em estrela pode ser visto na figura 5.

**Figura 5 - Modelo em Estrela**




Este modelo é composto por quatro tabelas de dimensão e uma de factos. Pretende representar as vendas de produtos postais nas estações de correios aos clientes e por tempo. Neste sentido, a tabela de dimensão *Produto* contém os produtos postais disponíveis nos Correios, a tabela de dimensão *Estação* contém as estações dos Correios, a tabela dimensão *Cliente* contém os clientes dos Correios e a tabela dimensão *Tempo* contém as datas (hora, dia, mês, ano). Quanto à tabela factos *Vendas* contém os dados relativos às vendas, por dia, produto, estação, cliente, quantidades vendidas, valor facturado, entre outros. De seguida ilustram-se, com um exemplo, estes conceitos. A tabela 1 mostra as vendas por estação.

**Tabela 1 - Vendas por estação**

| Estação | Vendas |
|---|---|
| Estação A | 5000 Euros |
| Estação B | 7000 Euros |

Adicionando a dimensão data (semestre), é aumentado o detalhe da análise (processo drill-down) que pode ser visto na tabela 2.

**Tabela 2 - Vendas por estação e semestre**

| Estação | Data | Vendas |
|---|---|---|
| Estação A | 1º Semestre | 3000 Euros |
| | 2º Semestre | 2000 Euros |
| Estação B | 1º Semestre | 4200 Euros |
| | 2º Semestre | 2800 Euros |

Aumentado ainda mais o detalhe, com a inclusão da dimensão *Produto*, obtêm-se a tabela 3:

**Tabela 3 - Vendas por estação, semestre e produto**

| Estação | Data | Produto | Vendas |
|---|---|---|---|
| Estação A | 1º Semestre | Correio Urgente | 1500 Euros |
| | | Correio Registado | 1000 Euros |
| | | Encomendas | 500 Euros |
| | 2º Semestre | Correio Urgente | 1100 Euros |
| | | Correio Registado | 700 Euros |
| | | Encomendas | 300 Euros |
| Estação B | 1º Semestre | Correio Urgente | 2600 Euros |
| | | Correio Registado | 1000 Euros |
| | | Encomendas | 600 Euros |
| | 2º Semestre | Correio Urgente | 1600 Euros |
| | | Correio Registado | 700 Euros |
| | | Encomendas | 500 Euros |

Poderíamos aumentar ainda mais o detalhe adicionando a dimensão *Cliente*, ou detalhar mais a dimensão *Data*, *descendo* ao mês ou ao dia. Contudo, para o exemplo pretendido, não nos parece necessário.

### 3.1.3. Data Mart

Quando se fala em DW, surge um outro termo, o Data Mart. Um Data Mart é em tudo idêntico a um DW, com a diferença que o Data Mart tem dados de uma área funcional e um DW tem dados de toda a organização.

Tipos de Data Marts

Os Data Marts podem ser dependentes ou independentes. É dependente quando detém os dados, relativos a um assunto, a partir de um DW. É independente quando os dados

vêm directamente das fontes de dados internas e/ou externas à organização, sem passar pelo DW.

Abordagem Top Down e Bottom Up

Estes dois tipos de Data Marts levam a dois tipos de abordagens: A abordagem Top Down e a Bottom Up. A abordagem Top Down tem como fundamento construir primeiro o DW e depois os vários Data Marts dependentes, alimentados com dados do DW. A abordagem Bottom Up é precisamente o contrário, ou seja, construir vários Data Marts independentes e alimentar o DW a partir desses Data Marts (Hackney, 1998).

Políticas na construção de um Data Mart

Segundo Atre (1997), os gestores acreditam muitas vezes que é mais fácil evitar as barreiras políticas e técnicas e construir Data Marts independentes. Custam menos inicialmente e podem necessitar apenas de aprovação departamental. Watson (1998) afirma que de facto muitas organizações optam pelos Data Marts independentes, pois o seu desenvolvimento é mais rápido com um menor custo e com grande retorno de investimento. Este autor afirma ainda que a construção deste tipo de Data Mart pode ser vista como uma "*prova de conceito*" para a construção do DW.

Contudo, defendemos que a prática deverá passar pela construção de um DW e, de seguida, construírem-se Data Marts departamentais, ou seja, Data Marts dependentes, alimentados a partir do DW. Infelizmente muitas organizações optam pelo contrário.

Percebemos que a implementação de um DW, embora possa ocorrer por fases, demora

sempre muito mais tempo que a construção de um Data Mart. Mesmo assim, os custos no final são bastante superiores pois existe a duplicação de esforços no processo de ETL, que consome aproximadamente 60% do tempo de implementação do projecto, uma vez que, enquanto numa abordagem Top Down o processo é efectuado apenas uma vez, numa abordagem Bottom Up o processo de ETL é efectuado tantas vezes quantos os Data Marts a construir. A abordagem Bottom Up tem ainda a desvantagem de proporcionar a criação de "ilhas" departamentais, e neste cenário, cada departamento apropria-se dos dados podendo dificultar o seu acesso a outros departamentos.

### 3.1.4. OLAP

O ambiente OLAP aparece devido ao OLTP, ambiente dos sistemas operacionais, não apresentar as características necessárias em termos de processamento de grandes quantidades de dados em tempo exigidas nos ambientes decisionais.

Características

O ambiente OLAP permite uma visão multidimensional dos dados, ou seja, permite uma visão dos negócios da organização em diferentes perspectivas. Para Codd (1970), OLAP é o nome dado à análise dinâmica necessária para criar, manipular, animar e sintetizar informação com modelos de análise de dados. A capacidade de visão do negócio em várias perspectivas é devida à capacidade de visão multidimensional dos dados, à facilidade de manipulação dos dados, à simplicidade de cruzamento de dados e à rapidez de processamentos de cálculo intensivo.

Operações OLAP

O ambiente OLAP permite efectuar operações multidimensionais como o *Drill-down* e *Drill-up*, sendo a primeira, a possibilidade de aumentar o nível de detalhe, e a segunda, o contrário, partir do detalhe para o geral. Operações de *Roll-down* e *Roll-up* permitem, respectivamente, aumentar ou diminuir o nível de agregação dos dados. A operação *Pivot*, consiste na mudança de orientação dimensional de uma pesquisa. Pode ser a troca de linhas ou colunas, ou mover uma das dimensões da coluna para a linha. As operações de *Slice* e *Dice* consistem em mudar a ordem das dimensões, mudando desta forma a orientação segundo a qual os dados são visualizados. Todas estas operações facilitam os decisores na análise dos dados, pois permitem diversas visões dos mesmos dados, levando o decisor a navegar pelos dados até encontrar os dados que lhe permitam ajudar a obter informação.

Benefícios do OLAP

Apresenta como benefício, conseguir fundamentar as decisões estratégicas através de várias análises, o que permite uma maior eficácia dos gestores (acesso a dados e informação de gestão e controlo ou indicadores e sem dependência de terceiros). É flexível, pois permite que os gestores façam previsões do estado do negócio, tendo como base os dados reais actualizados. Acompanha a evolução do negócio, permitindo que os gestores sigam os resultados das decisões tomadas e intervenham rapidamente sempre que existam desvios ao planeado.

Versões do OLAP

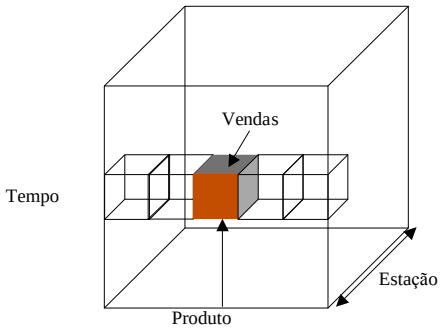
O MOLAP (Multidimensional OLAP), é um ambiente com BD proprietária onde não existe um modelo conceptual para representar os dados nesse tipo de SGBD. Mais tarde

surgiu outro ambiente denominado ROLAP (Relacional OLAP), ferramentas com BD relacionais, mas com os dados armazenados de forma desnormalizada. Segundo Donald (1997), este método foi implementado com muito sucesso, devido à facilidade de mapeamento, por utilizar o modelo relacional como base para a análise, e dispensar a utilização de uma base de dados multidimensional proprietária. Outro ambiente derivado do OLAP é o HOLAP, um híbrido entre o ROLAP e o MOLAP.

Exemplo de um cubo OLAP

Retomando o exemplo do modelo multidimensional, a Tabela 3 pode ser representada na forma de um cubo OLAP, conforme pode se observa na figura 6.

**Figura 6 - Cubo OLAP**




Com a rotação deste cubo é possível visualizar os dados de várias perspectivas, permitindo efectuar várias operações com os dados. Com a possibilidade de visualizar os dados em diferentes perspectivas consegue-se, neste caso, visualizar as *Vendas*, por dia, mês, trimestre, semestre, ano, para cada *Estação* e ainda por *Produto* postal. É ainda possível acrescentar mais dimensões como a dimensão *Cliente* e assim detalhar mais os dados.

Questões a que o cubo OLAP pode dar resposta

Algumas perguntas típicas que podem ser colocadas pelo decisor no ambiente OLAP: "Qual foi a variação percentual no total das vendas, em comparação com o mesmo período do ano anterior, para cada um dos 10 principais produtos, para cada um dos 10 maiores clientes?", "Quais são os dez produtos mais vendidos?", "Qual a média de vendas dos últimos três meses?", "Como se comparam as vendas deste ano com as do período homólogo?". É de realçar que nestas consultas os cálculos ocorrem simultaneamente em mais do que uma dimensão. Os cálculos do período anterior e do ano até à data actual são efectuados ao longo da dimensão *Data*. Ocorrem classificações (os dez mais) ao longo das dimensões *Cliente* e *Produto*, ou seja, a classificação do produto está aninhada dentro da classificação do Cliente.

### 3.1.5. Data Mining

O Data Mining é um processo que analisa os dados e indica, por exemplo, uma tendência. Esta ferramenta utiliza várias técnicas desde análise estatística, inteligência artificial, redes neuronais, algoritmos genéricos, reconhecimento de padrões, conjuntos difusos (Berson, 1997). Também podem ser aqui incluídas árvores de decisão e séries temporais. O Data Mining é o processo de descoberta de nova e relevante informação a partir de grandes volumes de dados mantidos no DW. Esses dados muitas vezes correspondem a "*conhecimento escondido*", que pode ser descoberto através da análise de padrões, bem como de correlações dos dados originais (Fayyad ***et al.***, 1996). O Data Mining é para os peritos e fornece as regras e relações.

### 3.1.6. As ferramentas de ETL

Uma ferramenta de Extracção, Transformação e Carregamento de dados, mais

conhecida por ferramenta de ETL, tem como principal objectivo tornar mais fácil a migração dos dados das BD operacionais ou de Sistemas Legados para o DW. É um processo complexo que consome tempo mas também é a base para assegurar um DW funcional (Vassiliadis *et al*., 2002). Estas ferramentas facilitam a vida a quem tem a tarefa de manipular dados. A QD no DW em muito depende da forma como os dados são tratados pelas ferramentas de ETL.

Requisitos

As ferramentas de ETL têm como principais requisitos o acesso aos dados, o controlo de processos de escalonamento, a flexibilidade e a forma como lidam com os metadados (Ballard e Herreman, 1998). No acesso aos dados, salienta-se a capacidade atribuída a estas ferramentas para extrair e carregar os dados de e em múltiplos repositórios, independentemente da tecnologia em que estes estão assentes. Estas ferramentas devem ser intuitivas, suportar funções de transformação, validação, tradução, integração, cálculo, agregação e derivação (operações que ocorrem na chamada área de estágio e que são detalhadas no capítulo 4). No controlo de processos e escalonamento devem ter a capacidade de suportar lançamentos automáticos, permitir a monitorização, a geração de relatórios e incluir módulos de segurança, de forma a controlar os acessos.

Principais operações das ferramentas de ETL

*Extracção*, processo associado à fonte de dados (BD operacionais e fontes externas), tendo como função extrair os dados destas fontes (Chuck *et al*., 1998). Deve causar o mínimo impacto nos Sistemas Operacionais. Há que tomar em consideração o esforço desta operação, cerca de 60% do tempo de construção do DW, principalmente se os

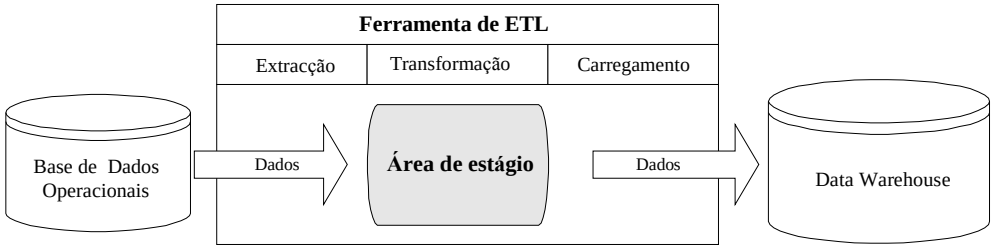
sistemas fonte forem antigos (mainframes), ou se os dados se encontrarem armazenados de uma forma desconhecida (Kimball, 1998).

*Transformação*, ocorre na área de estágio dos dados, e diz respeito à transformação que os dados estão sujeitos na passagem dos sistemas operacionais para o DW. Bohn (1997) defende que a QD no DW depende em muito de se fazer aqui um bom trabalho.

*Carregamento*, nesta operação são carregados e actualizados, no DW, os dados transformados que estão na área de estágio (Chuck *et al*., 1998). É ainda tida em conta a optimização da carga, o que pode levar à criação de índices.

As ferramentas de ETL, além de automatizarem a extracção, transformação e carregamento de dados, asseguram a actualização dos metadados associados no repositório de metadados (Ballard e Herreman, 1998). A figura abaixo, ilustra o fluxo de dados numa ferramenta de ETL.

**Figura 7 - Resumo do fluxo de dados**




### 3.1.7. Os Metadados

O termo metadado é definido como dados acerca de dados. Contêm dados de interesse para o utilizador sobre os dados que este usa, neste sentido, podem ser vistos como a *ficha técnica* dos dados. As pessoas ao usarem, no seu quotidiano, as tecnologias de informação servem-se de metadados sem terem consciência disso.

O propósito principal dos metadados consiste em dotar as organizações de informação para que estas sejam capazes de entender os seus dados e perceberem o que eles representam. Como consequência, os metadados reduzem a taxa de esforço e a facilidade de manutenção dos dados, e por conseguinte dos processos.

Os metadados fornecem informação detalhada, nomeadamente sobre a localização, a estrutura e a descrição dos dados. Descrevem ainda chaves e índices e fazem o mapeamento da informação com os algoritmos e com as regras de negócio usadas nos processos de tratamento e elaboração de agregações.

A grande quantidade e diversidade de dados existentes na Internet e nos sistemas de informação faz aumentar o interesse pelos metadados. Isto porque, os *utilizadores* ao pesquisarem os dados necessitam de proceder à sua validação, onde os metadados podem dar uma valiosa ajuda. Com o crescente uso de dados e número de utilizadores, a gestão dos metadados tornou-se fundamental.

Os metadados vão adquirir muito mais importância com o "*casamento*" da tecnologia da Web e do Data Warehousing. Esta união resultará num único ponto de acesso à informação do negócio, seja através de um "*Browser*", seja outro qualquer cliente como os Sistemas Operacionais ou o DW (SAS Institute, 2004). Os metadados tornam-se num dos componentes mais críticos para levar a cabo uma boa arquitectura informacional (Atre, 2003b).

Anteriormente aos anos 70 do século XX, os metadados estavam relacionados apenas

com os programas internos. Com o advento das Bases de Dados, ainda nos anos 70, as empresas enveredaram pela criação de Dicionários de Dados. Mais tarde surgiram as ferramentas CASE com os seus repositórios proprietários. Já nos anos 80, a IBM surgiu no mercado com uma proposta que consistia na criação de um repositório global para troca de metadados (AD/Cycle).

A integração dos sistemas, através dos metadados, tem sido um problema sem solução à vista. O rápido crescimento das organizações no que respeita a sistemas de informação tem tido como consequência a multiplicação de problemas com metadados. Estes problemas estão finalmente a ser admitidos pelos fabricantes de software, existindo neste momento algum trabalho desenvolvido no sentido de tornar compatíveis os metadados, ou melhorar os repositórios que os albergam, independente da tecnologia ou fabricante. Como consequência têm surgido no mercado várias propostas de normalização de modelos de metadados, como o uso de XML. De facto, os metadados são fundamentais para a compatibilidade dos sistemas. Por exemplo, ao adquirirmos um electrodoméstico preocupamo-nos com uma série de factores relacionados com as funcionalidades e características do aparelho, mas sabemos que ao optarmos por qualquer um, independente do fabricante, ele irá funcionar quando ligado à corrente eléctrica, isto porque, segundo a norma, todos funcionam ligados à corrente de 220 Volts.

Normalmente os sistemas de informação das organizações não são centralizados, o que coloca problemas de compatibilidade. Grande parte dos sistemas são desenvolvidos e instalados de acordo com uma visão local, com a finalidade de resolver um problema

singular. Embora esta abordagem não esteja propriamente incorrecta, ela pode dar origem a problemas quando houver necessidade de cruzar dados das várias aplicações. A integração desses dados é uma das tarefas e um desafio do DW.

Dados sem metadados podem ser vistos como uma corrida de Rally em que o co-piloto não tem o mapa da estrada ou do caminho, tornando-se quase impossível realizar a prova com sucesso uma vez faltar-lhe a informação necessária para perceber o caminho a percorrer, ou mesmo o caminho já percorrido.

O mesmo acontece com os dados, uma vez que, quando se analisa um determinado dado, ou conjunto de dados, torna-se necessário saber qual a sua origem, as suas regras de transformação e a sua finalidade. A obtenção dessa informação é possível graças aos metadados, ou seja, os metadados correspondem ao "*mapa*". Os metadados promovem a contextualização dos dados e, desta forma, ajudam a obter a informação pretendida.

Descrever o mundo real gera informação abstracta. Pode-se salientar o facto de ao descrever um fenómeno natural como a chuva, o vento ou o sol ser requerido um grande nível de abstracção. Qualquer destes três fenómenos está relacionado com o estado do tempo, logo com a temperatura, humidade, precipitação, etc. Conclui-se que existem vários níveis de detalhe, e como tal os metadados devem ser criados tendo em atenção estas questões, de forma a criar metadados com dados objectivos. É fundamental interpretar os metadados e perceber de forma clara e rápida toda a informação que eles possam gerar.

A existência de metadados relativos ao DW e ao DSS é importante para a compreensão do conteúdo e para a facilidade de manipulação dos dados. O modelo conceptual é independente da tecnologia usada e deste resulta o modelo lógico (modelo de dados). No modelo lógico são definidas as características dos dados a usar que serão armazenadas no repositório de metadados (RM), conhecido também por dicionário de dados ou catálogo.

Os metadados podem ser recolhidos de diferentes fontes tais como as ferramentas de modelação, os catálogos de bases de dados e as ferramentas de ETL. As organizações devem documentar os seus dados, caso contrário, com o decorrer do tempo, sujeitam-se a grande esforço de manutenção desses mesmos dados ficando vulneráveis a problemas de inconsistência com outros dados. Os metadados acompanham os dados durante toda a sua vida e vão sofrendo alterações ao longo dos processos de ETL a que estão sujeitos.

Papel dos metadados nas BD operacionais e no DW

Os metadados das BD operacionais têm um papel muito diferente dos metadados do DW. Nas BD operacionais podem realmente ser vistos como pura documentação, pois aqui o seu papel não é tão importante e, normalmente, não tem qualquer repositório associado (Inmon, 1997). A forma como os próprios dados são armazenados é menos complexa, já que existem poucas ou nenhumas agregações de dados. Mas nos projectos de DW, os metadados constituem a base para a integração dos modelos construídos durante o desenvolvimento do projecto, servindo de input à implementação da base de dados e dos códigos das aplicações (Moss e Atre, 2003). É importante ter uma boa documentação desde o início do projecto e, conforme refere Kimball e Caserta (2004),

nos projectos de DW os metadados são obrigatórios e quando incompletos ou pobres é sinónimo de má qualidade dos dados, e por conseguinte do DW.

Questões a que os metadados devem dar resposta

Os metadados devem responder a questões como:

- Que tabelas, atributos e chaves o Data Warehouse contém?
- Que dados estão disponíveis, em que área estão e quando foi a sua recolha?
- Quais foram as agregações criadas ?
- Que interrogações estão disponíveis para lhes aceder?
- Que pressupostos foram estabelecidos (quanto às regras de negócio)?
- Como aceder aos dados pretendidos?
- Qual a sua validade?
- Qual o volume de dados existente?
- O que significa ou representa um determinado conteúdo?

Estas são algumas questões a que os metadados têm de dar resposta uma vez que os dados apenas se podem transformar em informação se:

- Existirem.
- Soubermos que os temos.
- Soubermos onde estão.
- Pudermos aceder-lhes.
- Soubermos interpretá-los.
- Pudermos confiar neles.

Sem metadados, o utilizador não conseguiria interagir com os dados existentes num DW, ou noutro repositório, dada a impossibilidade de conhecer:

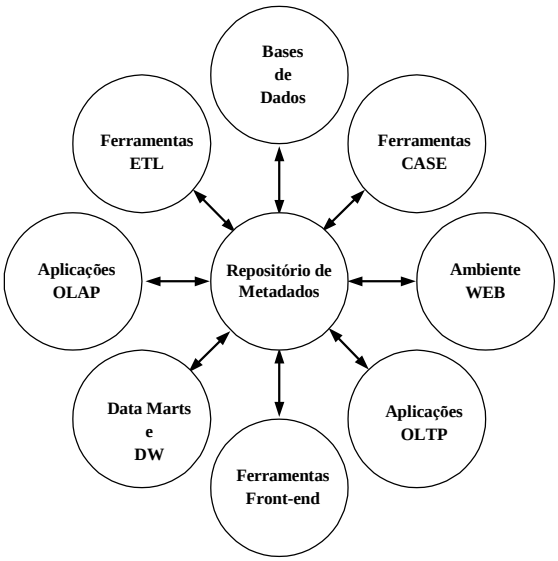
- O modelo de dados.
- A estrutura física.
- As definições dos dados.
- A origem dos dados.
- As datas de criação dos dados.
- O processo de extracção e transformação.
- As mudanças realizadas na estrutura ao longo do tempo.

A gestão dos metadados

A gestão dos metadados é vista como um dos factores críticos de sucesso dos projectos de Data Warehousing (Kimball, 1998). Não é fácil gerir os metadados, principalmente nos sistemas descentralizados e heterogéneos ou onde existam tecnologias diversas.

Segundo Satya (1998), a gestão de metadados deve ser aberta e disponibilizar a capacidade de manipulação dos metadados. Desta forma, qualquer ferramenta pode usar os metadados e criar outros dados para ajuda na obtenção de informação. A gestão de metadados deverá ter ainda como função facilitar a ligação de uma nova ferramenta bem como fornecer a capacidade de manipular diferentes fontes de metadados, como ilustra a Figura 8.

**Figura 8 - As fontes de metadados e o repositório de metadados**




Quando é necessário conceber alguns indicadores para a gestão é necessário saber onde estão os dados e como é possível aceder-lhes. Recorre-se a um conjunto de "*artifícios*" com o objectivo de recolher os dados e obter os indicadores. Mas colocam-se questões como: que indicadores gerar? Que credibilidade merecem? Qual a fonte dos dados?

Não é possível dar resposta a estas questões se não existir um repositório onde sejam registados os metadados. Caso não exista, apenas se pode contar com alguma documentação avulsa que exista sobre os dados fonte, e com o conhecimento da matéria que têm alguns colaboradores. Este é um ponto fraco nos sistemas de decisão, daí ser necessário armazenar os dados de forma centralizada e acessível a todos. É aqui que surge a nova abordagem à arquitectura de informação, e à cultura organizacional: é fundamental que, ao implementar-se novos sistemas de informação na organização, se proceda a uma actualização no RM. Mas para isso é necessário que ele exista e seja aceite por todos. Neste sentido, tem de haver por parte da organização uma cultura de comunicação e empenho para que todos vejam no RM uma forma de auxílio ao seu trabalho e passem a ter confiança nos dados que usam.

Problemas com a gestão dos metadados

Existem vários problemas relacionados com a questão dos metadados. Todos os dias podem surgir novos metadados, uma vez que a aquisição de novas tecnologias, desenvolvimento de novos processos e o meio produtivo diariamente geram ou podem gerar novos dados (novos dados na BD ou dados de monitorização de processos, como alertas ou relatórios de ocorrências). Conforme referem Moss e Atre (2003), nem todos os utilizadores vêm os metadados da mesma forma, pois dependendo da comunidade a que pertencem, os utilizadores podem usar diferentes vocabulários, termos e abordagens. Contudo, esta questão não pode afectar a centralização dos metadados. Assim, deve ter-se em atenção as normas, uma vez que podem comprometer a partilha dos sistemas (compatibilidade). Pode-se referir, a título de exemplo, os nomes usados, as versões, o controlo de acessos e a documentação.

Responsáveis pela gestão dos metadados

Os metadados são de tal importância que devem ser sujeitos a um controlo rigoroso. É necessário alguém responsável pela sua gestão e manutenção. A pessoa responsável deve assegurar que os metadados se mantêm actualizados e reflectem a infra-estrutura de negócio. Estes profissionais desempenham a função de *Administrador de Dados* (AD), o que incluí a definição e manutenção de conceitos e normas relativas ao(s) modelo(s) de dados/informação. O AD é responsável pelo modelo de dados, devendo interagir com a Gestão de Topo, Utilizadores e Equipas de Desenvolvimento Informático (ED). Segundo Date (1995), o AD é um gestor e não um técnico. Este ponto de vista é pertinente, pois um técnico pode não ter as capacidades de negócio adequadas para "*classificar*" os dados. Contudo, conforme sugere Satya (1998), este

gestor tem de ter uma componente técnica e estar bem informado quanto às tecnologias de informação disponíveis.

A função de *Administrador de Bases de Dados* (ABD) revela-se igualmente importante ao nível da gestão de dados, uma vez que é responsável pela implementação da estrutura dos dados, tarefa que deve ser realizada em conjunto com a administração de dados e em colaboração com as equipas de desenvolvimento. São ainda da sua responsabilidade a implementação de mecanismos de integridade, controlo de segurança e de recuperação das bases de dados, para além da monitorização do respectivo desempenho.

O conceito de metadados é essencial para o desenho e utilização de qualquer sistema de informação. Independentemente de qual a função, todos os intervenientes tem que, directa ou indirectamente, utilizar os metadados como forma de identificar campos, tabelas, fontes de informação e caminhos de acesso (Inmon e Hackathorn, 1994).

De acordo com Singh (1997), os profissionais que compõem as equipas de desenvolvimento podem vir de diferentes áreas de negócio tais como, análise de negócio, modelação de dados, administração de dados e de bases de dados, desenvolvimento de aplicações ou mesmo da área de apoio ao utilizador final.

Singh (1997) reforça ainda a ideia de que os metadados são uma das principais componente de um DW. Neste sentido, pensamos ser imprescindível o acesso aos metadados por parte do ABD, do AD e das ED. Devem estar disponíveis desde a identificação dos dados até à criação de pesquisas no portal. No fundo, devem

acompanhar todo o ambiente operacional, o ambiente de armazenamento e o ambiente analítico ou ainda todo o ciclo de vida dos dados. A presença dos metadados permite que os utilizadores determinem a qualidade das pesquisas e análises. Sem eles, os dados poderiam não ter significado, ou pelo menos a sua boa utilização podia ser posta em causa. O grau de qualidade de um sistema de decisão deve-se em muito à qualidade dos metadados, pois é através destes que os utilizadores confiam nos dados e na informação obtida.

Os metadados possibilitam a obtenção de informação sobre o nível de optimização do sistema (DW, Data Mart), e obter estatísticas sobre a utilização destes. O seu uso permite ao AD ou ao ABD efectuar a optimização, monitorização e ajuste da Base de Dados de maneira a melhorar o seu desempenho. Bem como fazer um tratamento estatístico prevendo um conjunto de acontecimentos. No caso do AD ou do ABD, pode ser possível prever o crescimento do volume de dados numa determinada tabela e aumentar o *espaço de armazenamento* com alguma antecedência evitando que a aplicação "aborte" por falta de espaço. Podem ser optimizadas as consultas ad-hoc, pois é possível aceder às pesquisas realizadas pelos utilizadores e sugerir outro método de fazer mais rápido. Existem ferramentas no mercado para ajudar os AD e os ABD nestas tarefas.

Tipos de metadados

De acordo com Brackett (1996), existem dois tipos de metadados: os técnicos, usados pelos ABD e Programadores, e os de negócio, usados pelos utilizadores finais.

Metadados Técnicos

São utilizados pelas equipas de desenvolvimento e manutenção com o objectivo de conhecer as fontes dos dados, as regras que foram utilizadas na sua criação e o seu significado. Com efeito, é através destes metadados que se pode obter uma ajuda preciosa na compreensão e controlo dos vários sistemas relativamente a aspectos como: as fonte dos dados, a frequência de actualização, as dependências recíprocas, a data/hora do último carregamento, as regras de negócio aplicadas aos dados dos sistemas fonte, que processos dependem do carregamento desta informação, o nome do processo que realizou o carregamento, os índices, os nomes de tabelas e as chaves primárias. David Marco (1998) considera ainda o balanceamento, o mapeamento, o modelo lógico dos dados, os nomes dos programas e as descrições dos dados como fazendo parte de metadados técnicos.

Metadados de Negócio

Permitem aos utilizadores interpretar os dados que lhes são disponibilizados em relatórios, gráficos ou através de consultas em ambientes analíticos ou interrogações ad-hoc. Os utilizadores que recorrem a estes metadados são, normalmente, executivos ou analistas de negócio. Estes utilizadores tendem a ser menos técnicos, logo necessitam que os metadados lhes forneçam informação como: que dados, relatórios ou consultas se encontram definidas no DW, qual a localização dos dados, qual a confiabilidade, o contexto, as regras de transformação que foram aplicadas e quais as origens desses dados. No fundo ajudam a perceber o contexto do negócio e o significado dos dados.

O Repositório de Metadados (RM)

A criação de um RM é especialmente crítica no momento de converter dados em informação para gerir o negócio. Segundo Jennings (2003), "*a integração entre os sistemas fonte operacionais, o Data Warehouse, os processos de extracção, transformação e carregamento (ETL), as regras de negócio, os relatórios e as estatísticas operacionais ocorre sobre o controlo do repositório de metadados*". O RM, muitas vezes designado por dicionário de dados, deve incluir, entre outras coisas, objectos de negócio e respectivos atributos, fontes de dados, regras de transformação, regras de negócio, processos, formas de acessos e regras de segurança (Lebaron e Adelman, 1997).

É de salientar o facto de o projecto de gestão de metadados estar mais focado nos problemas de integração de sistemas e dados do que no esforço de desenvolvimento da própria ferramenta de gestão de metadados.

A diversidade de ferramentas existentes nos sistemas de informação, especialmente nos sistemas de suporte à decisão, tem sido uma das maiores dificuldades no desenvolvimento de um RM de qualidade, levando os fabricantes de software a juntar esforços no desenvolvimento de protocolos e tecnologias onde seja possível a interacção com os metadados (Booch *et al*., 1999).

Componentes básicos do RM

Mike Jennings (2003), enumera ainda os sete componentes básicos do RM:

1) Modelo lógico do *Data Warehouse.*

2) Modelo físico do *Data Warehouse.*

3) Modelos de dados dos sistemas fonte.

4) Regras de mapeamento dos dados fonte para os de destino.

5) Áreas de negócio.

6) Estatísticas dos processos de extracção, limpeza e carregamento (*ETL*).

7) Estatísticas das consultas feitas ao *Data Warehouse.*

Funcionamento do RM

Existem ferramentas que facilitam a consulta e manipulação no RM bem como a identificação dos metadados associados às funções de captura e integração, manipulação e gestão do DW (Sherman, 1997; Seligman e Rosenthal, 1996). O RM vai abarcar todos os metadados resultantes dos processos de ETL, ou seja, da extracção dos dados do ambiente de produção (OLTP) para o ambiente do DW (OLAP), tendo em conta que pelo meio são gerados novos metadados respeitantes a processos de transformação, validação e ainda metadados resultantes de carregamentos dos dados.

No processo de transformação dos dados podem ocorrer inconsistências de formatos. Tendo em conta que as regras de transformação são estabelecidas pelo *Administrador de Dados*, este pode igualmente criar e incluir novas regras. Conforme salientam Ballard e Herreman (1998), estas transformações são armazenadas como metadados. Assim, para criar uma arquitectura de metadados e, por conseguinte, geri-los, devem-se ter em conta as seguintes tarefas:

- Definir os requisitos.
- Desenvolver uma arquitectura de gestão.
- Seleccionar as ferramentas de gestão de metadados.
- Desenvolver programas que integrem e adaptem as ferramentas seleccionadas

para atender às necessidades de gestão.

- Desenvolver um plano de formação para os utilizadores que vão usar este ambiente de gestão.

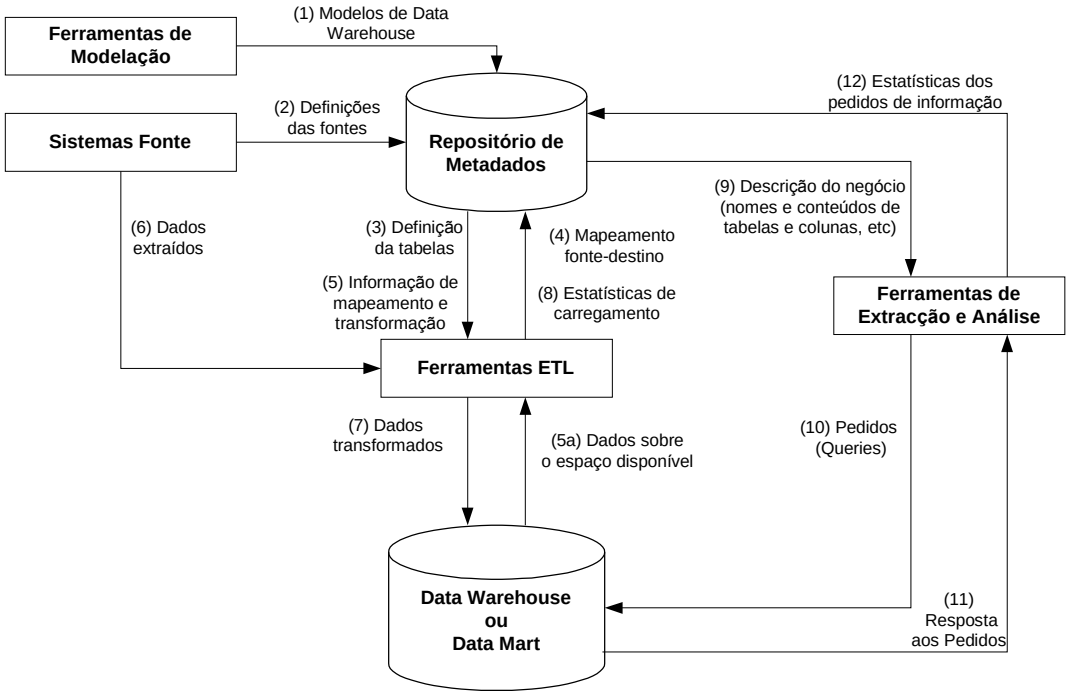
Benefícios do RM

Como principal benefício do RM pode-se salientar o facto de permitir uma plataforma de integração e armazenamento de metadados a toda a organização. Desta forma, torna possível a manutenção dos metadados e possibilita, entre os utilizadores, a partilha das estruturas de dados comuns, as definições de regras de negócio e as definições de dados dos diferentes sistemas da organização. É de salientar ainda o papel activo que o RM tem em todo o ambiente organizacional, desde o carregamento dos dados dos sistemas táctico-operacionais até à disponibilização ao decisor através das ferramentas de BI.

Exemplo de possíveis fluxos de dados que interagem com o RM

A figura 9 pretende ilustrar possíveis fluxos de dados de alguns dos componentes que interagem com o RM.

**Figura 9 - Fluxo de dados que interagem com o repositório de metadados**




O *primeiro* passo consiste na introdução das definições dos modelos no RM através das ferramentas de modelação. Esta informação inclui os nomes das colunas, nomes físicos das colunas, conceitos e descrições de negócio associados e valores de exemplo. Em seguida, *segundo* passo, capturam-se as definições dos dados fonte onde, se pode utilizar uma ferramenta ETL já que depende muito da informação gerada por esta ferramenta. O *terceiro* passo consiste na captura das definições das tabelas de destino. O *quarto* passo consta na definição das regras de mapeamento entre as fontes e os destinos. Também aqui é potenciado o uso das ferramentas ETL. No passo seguinte, o *quinto*, interrogam-se os metadados de forma a saber tudo sobre as fontes, os destinos e as transformações dos dados de forma a carregar os dados no DW ou Data Mart. Pode igualmente interrogar-se a base de dados de destino de forma a obter informação sobre o estado físico do sistema. No *sexto* passo extraem-se os dados fonte, que são transformados e carregados no DW ou Data Mart no *sétimo* passo. Seguidamente, no

passo *oitavo*, são capturadas algumas estatísticas acerca do carregamento e guardadas no RM. Os termos específicos de negócio existentes no catálogo de metadados são, então, utilizadas por uma ferramenta de extracção de dados para que o utilizador não necessite de saber os nomes das colunas da base de dados (*nono* passo). No *décimo* passo, o utilizador efectua o pedido de informação que pretende consultar. Aqui igualmente são usados os dados contidos no RM para uma construção optimizada do comando. No *décimo primeiro* passo são devolvidos os resultados e, no último passo (*décimo segundo*), são guardados os dados estatísticos sobre esta operação de modo a que os processos possam ser analisados e optimizados.

# 4. A Qualidade dos Dados - Dados no Data Warehouse e Dados disponibilizados pelas ferramentas de BI

## 4.1. Elementos introdutórios sobre qualidade

No cenário competitivo em que vivemos é importante desenvolver produtos e prestar serviços com qualidade, sem trabalho redobrado e de maneira inovadora. É fundamental aliar o método (forma de fazer), as ferramentas de qualidade, a criatividade e a inovação para que as organizações se possam diferenciar neste mercado. Garantir a qualidade não é tarefa fácil uma vez que cada um dos intervenientes tem, ou pode ter, pontos de vista diferentes quanto ao que entende ser a qualidade. Na sequência desta dissertação entende-se ser pertinente apresentar algumas definições de qualidade, organizadas cronologicamente na tabela 4.

**Tabela 4 - Definições da Qualidade**

| Autor | Definição de qualidade |
|---|---|
| Jenkins (1971) | "*grau de ajuste de um produto à procura que pretende satisfazer*" |
| A OECQ (1972) | "*aptidão para o fim a que se destina*" |
| Juran (1974) | "*aptidão ao uso*", "*a qualidade é a ausência de falhas no produto que, através das suas características, satisfaz o cliente e vai ao encontro das suas necessidades*" |
| Crosby (1979) | "*conformidade com as especificações*"; "*conformidade com os requisitos"; "é prevenção"; "zero defeitos";* |
| Taguchi e Wu (1979) | "*perda para a sociedade, causada pelo produto ou serviço, após o seu fornecimento ou expedição*" |
| Kano (1984) | *"as expectativas dos clientes podem ter que ser excedidas. As necessidades básicas e as experiências excitantes"* |
| Norma ISO 8420 (1994) | *"...totalidade das características de um produto, processo ou serviço, que suportam a sua capacidade de satisfazer necessidades explícitas e implícitas"* |
| Goetsch e Davis (1997) | *"estado dinâmico associado a produtos, serviços, pessoas, processos e meios envolventes que vão ao encontro ou excedem expectativas"* |
| Paladini (2000) | *"Possui uma componente espacial, a multiplicidade de itens, e uma componente temporal, as alterações conceptuais ao longo do tempo (processo evolutivo) "* |

A qualidade pode ser definida como o grau de perfeição a atingir, bem como a melhor forma de atender às necessidades dos utilizadores, tendo em conta a finalidade do produto. Assim, a qualidade pode ser vista como o conjunto de atributos que tornam um bem ou serviço plenamente adequado ao uso para o qual foi concebido. Conforme refere Sousa-Mendes (2001a), "*qualidade é a capacidade de algo cumprir perfeitamente os objectivos para que foi concebido*".

Apesar da inerente subjectividade, a qualidade de um produto ou serviço depende sempre de vários factores – a finalidade, os equipamentos, os materiais ou os métodos empregados – bem como de diversos critérios - a operacionalidade, a segurança, a tolerância a falhas, o conforto, a durabilidade e a facilidade de manutenção.

A qualidade de um produto é decorrente da qualidade do processo de produção e para se obter um produto com qualidade é necessário acompanhar o seu *ciclo de vida*. A qualidade é ainda o resultado de um esforço no sentido de desenvolver um produto ou serviço de modo tal que este atenda a determinadas especificações. A qualidade "***absoluta***" não existe e não se consegue atingir a qualidade se esta não for especificada.

A qualidade é a satisfação total do cliente, e esta satisfação é reforçada pelo facto das organizações serem confrontadas com uma nova realidade. Por um lado, o aumento da oferta em relação à procura e, por outro, o aumento da capacidade de escolha do cliente. O surgimento de movimentos e organizações de defesa do consumidor, como por exemplo a DECO, também contribuíram de forma preponderante para a importância dada à qualidade, seja pelas organizações, seja pelos consumidores.

Ao longo da história o entendimento da qualidade tem sofrido evoluções, todas elas originadas pelas forças de mercado. Há que realçar o facto de a qualidade existir porque existem consumidores. Certo é, que a maior ou menor qualidade também depende do grau de exigência dos consumidores e da competitividade dos mercados, sendo que esta última pode ser vista como uma forma de obter vantagens competitivas.

### 4.1.1. Evolução da qualidade

A preocupação com a qualidade do produto não é apanágio da sociedade competitiva em que vivemos. Mesmo nas sociedades mais remotas, esta preocupação estava presente. Inicialmente, o artesão, único interveniente no ciclo de vida do produto, desempenhava várias funções, entre as quais a identificação das necessidades, a concepção do produto, a sua fabricação, a comercialização e os serviços pós-venda (Pires, 2000).

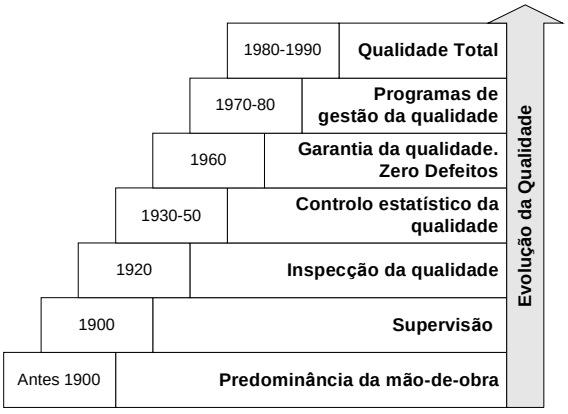
A necessidade de aumentar a produção teve como consequência o aparecimento e aumento do número de oficinas, espaços onde a hierarquia se encontrava devidamente definida e onde a divisão de tarefas era já uma realidade: o mestre, que assegurava a gestão e direcção da oficina bem como dos acabamentos de maior precisão, o ajudante, aquele a quem o mestre delegava o papel de controlo do aprendiz, último na hierarquia. O produto final resulta da acção destes intervenientes, sendo que a qualidade era assegurada pelo ***saber-fazer*** do mestre.

Pode-se concluir que antes da produção em série, o homem (ou grupo) era responsável pela qualidade do seu próprio trabalho (predominância da *mão-de-obra*). O controlo da qualidade pelo operador foi uma realidade até aos fins do século XIX.

No início do século XX, verifica-se a primeira transformação, uma vez que a qualidade passa a centrar-se no campo da *supervisão.* Por volta dos anos 20 surge o conceito de *inspecção* da qualidade do produto (verificação do respeito das especificações internas em vigor) e nas décadas de 30-50 o processo volta a evoluir e a garantia da qualidade encaminha-se para o *controlo estatístico e fiabilidade* (identificação do problema de produção).

Nos anos 60 começa a falar-se em *garantia da qualidade, motivação para a qualidade e zero defeitos* (prevenção do problema e não a sua detecção). Contudo, nos anos 70 e 80 surge uma nova abordagem através dos *Programas de Gestão da Qualidade*, também conhecidos por círculos da qualidade, caracterizados por uma competitividade baseada em estratégias quantitativas que procuravam satisfazer os mercados em expansão, onde a procura era maior que a oferta. Nas décadas de 80 e 90 surge a chamada *qualidade total* e a qualidade gradualmente torna-se o factor mais importante da competitividade.

**Figura 10 -** Evolução do conceito de qualidade[2]




[2] Esta figura tem como base a apresentada por Pires (2000, p.34)

## 4.2. Os Dados em ambiente OLTP vs. OLAP

Os dados no ambiente OLTP

A maioria dos dados residentes no DW tem origem nas BD operacionais, e em ambos os sistemas é necessária a garantia da qualidade dos dados. Contudo, as dimensões de QD nas BD operacionais não são necessariamente as mesmas que no DW (Inmon, 1997). Os Sistemas Operacionais, conforme o nome sugere, são sistemas de apoio às operações de negócio, e Singh (1997) afirma que o objectivo destes é essencialmente operar o negócio, devendo a disponibilidade dos dados ser total, isto é, a consulta ou recolha de dados deve ser um processo fácil e rápido. Verifica-se que o uso diário dos dados pelos colaboradores no desempenho das suas funções não exige que estes tenham grandes conhecimentos de Sistemas de Informação, uma vez que as consultas que fazem são limitadas e baseadas num pequeno número de dados.

Os dados no ambiente OLAP

Em contrapartida, o uso do DW exige colaboradores com conhecimentos específicos. Segundo Berson (1997), estes trabalhadores manipulam grandes quantidades de dados com o objectivo de apoiar os decisores aos níveis táctico e estratégico, sendo que os dados disponibilizados têm como objectivo analisar o negócio. Assim, o DW é um sistema específico, uma vez que cada organização tem as suas necessidades de negócio.

Características dos dados nos dois ambientes

Segundo Kimball (1998), os Sistemas Operacionais podem processar milhares ou milhões de transações por dia usando uma pequena quantidade de dados, enquanto que o DW processa poucas transações por dia mas usa milhares ou milhões de dados.

Na tabela 5 apresenta-se um resumo das diferentes características dos dados nos dois ambientes.

**Tabela 5 - Características dos dados nos ambientes OLTP vs. OLAP**

| Características dos dados | Ambiente OLTP | Ambiente OLAP |
|---|---|---|
| Tipo | Detalhados<br>Actuais e voláteis | Detalhados e sumariados<br>Históricos e não voláteis |
| Organização | Por aplicação | Por assunto |
| Estabilidade | Dinâmicos | Estáticos |
| Optimização | Para transações | Para pesquisas complexas |
| Dados por transação | Poucos (dezenas) | Muitos (milhares) |
| Frequência de acesso | Alta | Média ou baixa |
| Volume de dados | Megabytes/Gigabytes | Gigabytes/Terabytes |
| Tipos de operações | Actualização e consultas | Consulta e análise |
| Processamento | Focados na transação | Focados na análise do negócio |
| Uso | Dirigidos às operações | Dirigidos à análise estratégica |
| Área de negócio | Funcional e Operacional<br>(decisões no dia-a-dia) | Estratégica<br>(Decisões no longo prazo) |
| Redundância | Controlada | Obrigatória |
| Interacção | Pré-definida | Pré-definida e ad-hoc |
| Actualização | Em tempo real | Periódica (operações batch) |
| Disponibilidade | Alta | Atenuada |
| Modelação | Entidade-Relacionamento | Multidimensional |

## 4.3. A Qualidade dos Dados no Ambiente Analítico

Para que se inicie um processo de melhoria de QD nas organizações estas devem procurar gerir correctamente os seus recursos informacionais. Para isso, devem atender, nomeadamente, à identificação de *Dados* e *Informação* relevantes para as suas operações, ao desenvolvimento de práticas que assegurem a Qualidade dos Dados e da Informação e proceder à disponibilização de tais recursos aos departamentos que deles necessitem (Gartner Group, 2004).

Presentemente, as grandes vantagens competitivas nas organizações podem advir, em parte, dos dados, se a qualidade destes for uma realidade. Como refere Berson (1997), em geral, a falta de exactidão dos dados, o facto de estarem incompletos ou serem antiquados tem um impacto social e económico significativo nas organizações.

Para Ken Orr (1998), gerir a QD é uma tarefa complicada e complexa mas nem sempre é necessário chegar ao limite de "*zero defeitos*". Em primeiro lugar, porque não é imperativo que algumas aplicações usem dados com "*zero defeitos*", e em segundo, porque existem custos inerentes, quase sempre elevados, para atingir os "*zero defeitos*".

Como consequência, torna-se necessário perceber até onde se quer ir na gestão da QD e cabe a esta gestão delinear a melhor estratégia a seguir. Para isso, deve considerar factores como os custos, o tempo despendido e o benefício obtido. Ballou e Pazer (1987) afirmam que, na maioria dos casos, a melhor solução em termos de redução da percentagem de erros pode ser a pior em termos de custo.

Ken Orr (1998), afirma que nenhum sistema de informação tem QD a 100%. Este autor refere ainda que a QD não é assegurar que os dados sejam perfeitos mas sim que a sua qualidade possa assegurar a sobrevivência da organização e possa ajudar a tomar decisões sensatas. Ainda na linha da melhoria da QD, Redman (1998) afirma que a falta de QD gera um impacto bastante negativo nas organizações, até porque os custos operacionais podem aumentar, e a confiança dos trabalhadores e dos consumidores diminuir.

É necessário analisar e melhorar a QD, e para o efeito podem ser usadas ferramentas de controlo de qualidade. Com o contributo destas é possível identificar os factores que mais contribuem para a falta de qualidade sendo que, por vezes, melhorando um conjunto mínimo de factores se consegue aumentar significativamente a QD. A falta de qualidade dos dados corresponde a uma pobre qualidade da informação, podendo

resultar numa análise deficiente e, subsequentemente, numa má decisão. E uma má decisão no ambiente económico dos nossos dias, conforme referido no capítulo 2, poderá ter consequências terríveis para as organizações.

Da mesma forma que é difícil gerir a qualidade dos produtos sem compreender quais são as características que a definem, também se torna difícil gerir a qualidade dos dados sem perceber as suas características. Para isso, verifica-se a necessidade de perceber quais as dimensões de QD, tendo como suporte a pesquisa e interpretação de bibliografia especializada. Após esta fase de análise e interpretação propõe-se um conjunto de dimensões de QD no DW e nas ferramentas de BI.

#### 4.3.1.1. Pesquisa bibliográfica

Na pesquisada efectuada encontraram-se diversas propostas de dimensões de QD e neste capítulo pretende-se dar a conhecer tais propostas. Com vista a uma ilustração do pensamento dos autores e da construção do próprio pensamento relativo ao tema, as referidas propostas encontram-se organizadas cronologicamente.

Morey (1982), considera a dimensão ***Precisão*** como das mais importantes para a QD e refere que esta dimensão tem que ver com o facto do valor registado estar ou não em conformidade com o valor actual. Aponta ainda como factor que contribui para a falta de QD a ocorrência de erros, relacionada com as demoras de processamento dos dados, a demora na correcção ou a incompreensão dos dados.

Por sua vez, Ballou e Pazer (1987) apontam como principais dimensões a ***Oportunidade***, o valor registado que não está obsoleto, a ***Completude***[3], a situação em que todos os valores de uma determinada variável estão registados, e a ***Consistência***, a representação dos dados é a mesma, independentemente do ambiente.

Já na década de 90, Huh *et al*. (1990) consideram ***Precisão***, ***Perfeição***, ***Consistência*** *e* ***Correcção*** como as dimensões mais importantes e as que asseguram a QD. Por sua vez, Wand e Wang (1996), num artigo publicado em Novembro na revista *Communications of the ACM*, fazem referência à investigação levada a cabo por Wang, Storey e Firth (1995), baseada na análise de artigos e citações relativos à área da QD. Nesse estudo procura-se perceber quais as dimensões mais citadas nos documentos e concluem que são a ***Perfeição***, a ***Ambiguidade***, a ***Relevância*** e a ***Correcção***, sendo também estas as dimensões que Wand e Wang (1996) sugerem como as mais importantes.

No mesmo ano, Wang e Strong (1996) propõem uma framework, baseada em mais de 150 requisitos, com a finalidade de analisar as dimensões da QD, isto porque, no entender destes autores, a QD pode ter diferentes significados para diferentes utilizadores. De facto, a qualidade não é sentida da mesma forma por todos os utilizadores, variando igualmente de acordo com o tipo de ambiente, seja este OLTP ou OLAP. Tendo em conta a subjectividade da definição de QD, os referidos autores propõem-se desenvolver nessa framework um mecanismo para que todos os utilizadores identifiquem as dimensões de QD mais apropriadas para as aplicações que usam.

[3] Este termo "completude" foi retirado da tradução feita por Sousa-Mendes (2001a) na sua dissertação, *A Qualidade dos Dados nos Sistemas de Informação*, onde procura dar sentido à palavra inglesa "completeness".

Para Thomas Redman (1996), as dimensões da QD devem ser entendidas em três perpectivas: a Conceptual, que integra as dimensões de *Detalhe*, *Consistência*, *Composição*, *Robustez* e *Flexibilidade*; o Valor dos Dados, com a *Precisão*, a *Perfeição*, a *Correcção* e a *Consistência*; e, como última perspectiva, a Representação dos Dados, com as dimensões de *Apropriação*, *Interpretabilidade* e *Portabilidade*.

Dois anos mais tarde, Richard Wang (1998) sugere a divisão das dimensões da QD nas categorias representadas na tabela 6:

**Tabela 6 - Proposta de divisão das dimensões**

| Categorias | Dimensões |
|---|---|
| Intrínseca | *Precisão, Objectividade, Confiabilidade, Reputação* |
| Acessível | *Acessibilidade, Segurança* |
| Contextual | *Relevância, Oportunidade, Perfeição, Adequação ao negócio* |
| Representativa | *Interpretatibilidade, Correcção, Consistência, Endereçabilidade* |

Mais recentemente, Pipino, Lee e Wang (2002), no artigo *Data Quality Assessment*, publicado na revista *Communications of the ACM*, apresentam um conjunto de dimensões, baseadas num questionário realizado para determinar a percepção da QD pelos intervenientes. As dimensões consideradas mais importantes neste estudo são: *Acessibilidade*, *Interpretatibilidade*, *Apropriação*, *Confiabilidade*, *Correcção*, *Perfeição*, *Objectividade*, *Reputação*, Segurança, *Adequação ao negócio* e *Compreensão*.

Também o Meta Group, mais propriamente Goggin (2003), identifica nos seus estudos um conjunto de dimensões para garantir a QD, são elas, a *Precisão*, a *Perfeição*, a *Consistência*, a *Correcção*, a *Endereçabilidade*, a *Oportunidade*, a *Redundância* e a *Integridade*.

#### 4.3.1.2. Comentários à bibliografia pesquisada

Após apresentação das dimensões de QD propostas pelos autores, constata-se que existem dimensões mais direccionadas para a qualidade dos dados, outras para a qualidade da informação e ainda outras para a qualidade dos sistemas. Verifica-se ainda que os autores não separam as dimensões de QD das BD operacionais das dimensões de QD do DW.

Pensamos ser correcto afirmar que, quando se iniciaram os processos de migração de dados para o DW não foi tida em conta a QD, pois considerou-se que estes já tinham qualidade. Contudo, os ambientes OLTP e OLAP são diferentes e têm objectivos distintos, neste sentido, as Dimensões de QD não coincidem na íntegra. Ao migrar os dados das diferentes plataformas para um DW, a QD geralmente é afectada como tal, deve-se usar uma ferramenta de ETL que apoie este processo e ajude a garantir a QD no DW.

Partindo do princípio que as dimensões de QD para as BD operacionais estão identificadas, e são as propostas por Sousa-Mendes (2001a), verifica-se a necessidade de identificar e propor um conjunto de dimensões de QD para o DW e também para as ferramentas de BI.

#### 4.3.1.3. Proposta de Dimensões de QD no Data Warehouse

Propomos de seguida, baseado na bibliografia pesquisada, um conjunto de dimensões de QD para o Data Warehouse.

*Precisão*: tem a ver com o conteúdo e com o domínio. Permite detectar problemas como: valores muito fora do esperado (*Outliers* ou ocorrências negativas); incoerência entre o tamanho do campo e a documentação ou especificação; imprecisão que advém dos arredondamentos, principalmente quando se copiam dados de um local para outro; não identificação da escala usada (por exemplo perceber se estamos à espera de percentagens, unidades de medida, ou outras); incoerências de formatação (formatação da data, da hora ou mesmo dos códigos postais).

*Adequação ao negócio*: esta é uma dimensão que abrange todos os dados pois eles devem respeitar as regras de negócio. Esta dimensão ajuda a perceber como as entidades estão referenciadas na organização. Para uma melhor compreensão, há que identificar os sinónimos**,** isto é, palavras diferentes que podem ou não representar o mesmo (num_emp e n_emp é o nome do campo número de empregado). Também se deve ter especial atenção aos homógrafos**,** palavras que se escrevem de forma igual mas que representam coisas diferentes, dependendo do contexto.

*Correcção*: tem que ver com o conteúdo do dado e a sua fonte, isto porque, para que um determinado dado se possa considerar correcto tem de estar coincidente com a sua fonte. Nesta categoria analisa-se o impacto da transformação de dados do sistema de origem para o sistema de destino. Permite ainda avaliar se o processo de agrupamento de dados está correcto e analisar o mapeamento dos dados para que todos tenham correspondência no novo sistema.

*Relevância*: tem a ver com o grau de importância de um determinado dado para as análises que se pretende levar a cabo, isto é, se o dado é fulcral na tarefa para a qual é usado. Mais uma vez as regras de negócio ajudam a determinar se um dado é ou não relevante.

*Referenciabilidade*: tem a ver com o facto de o dado dever estar associado ou não a uma unidade de referência. Por exemplo – O campo quantidade com valor de 50 tem de estar referenciado por uma determinada unidade (metros, quilos, unidade monetária); só assim pode ser interpretado, caso contrário não se sabe a que se refere. Conforme refere Sousa-Mendes (2001a), um dado diz-se referenciável se existir um outro dado sem o qual não seja possível a sua plena interpretação.

*Oportunidade*: reflecte a disponibilidade dos dados em tempo útil e tem que ver com as regras de negócio, isto porque o sentido dado à oportunidade difere de negócio para negócio. Existem organizações que apenas necessitam de dados recentes, outras que necessitam de dados históricos e ainda algumas que carecem de ambos. Por exemplo – o resultado eleitoral ou a importação de um determinado produto nos anos 20, são dados que podem contribuir para o cálculo de um determinado indicador.

*Objectividade*: refere-se à imparcialidade e à independência com que os dados são migrados para posterior auxílio na obtenção de informação para a tomada de decisão. Nem todos os dados residentes nos Sistemas Operacionais migram para o DW, daí a necessidade dos trabalhadores do conhecimento que, conforme referido por English (1999), têm grande capacidade de operar na migração de dados. Este profissionais têm

grande percepção do negócio e assim conseguem maior percepção na identificação dos dados a passar para o DW. De nada vale ter dados que não são usados pois só carrega o DW com dados inúteis, afectando o desempenho do sistema.

*Desempenho*: tem que ver com o tempo de resposta de uma determinada consulta e, neste contexto, é importante o uso de dados agregados de forma a reduzir esse mesmo tempo (Winter, 1999). Justifica-se, assim, esta dimensão de qualidade, que terá como objectivo avaliar a capacidade de leitura dos dados (input/output) e a rapidez de cálculo dos mesmos.

*Tamanho*: no ambiente OLAP, o número de registos nalgumas tabelas é muito grande contudo, importa perceber se todos são necessários. Esta dimensão tem ainda como finalidade avaliar o número de tabelas de factos. Estas são muito importantes num modelo dimensional, uma vez que comportam dados de medição resultantes dos processos de negócio, pois quantas mais forem, maior será o espaço que ocupam e mais difícil a resposta às consultas. Segundo Inmon (1997), estas tabelas de factos ocupam cerca de 90% do espaço total do DW. Um DW tem várias tabelas de factos e cada uma relacionada com um número de tabelas de dimensões, que normalmente anda na ordem das 5 a 15. As tabelas de dimensões contêm descritores textuais do negócio, possuem bastantes atributos e não é difícil encontrar 50 ou 100 atributos nestas tabelas. Não obstante esta situação, é de todo importante perceber se as tabelas de dimensões não se repetem, ou seja, se as tabelas de dimensões podem ou não estar relacionadas com várias tabelas de factos. Por exemplo, a tabela de dimensão *tempo,* serve várias tabelas de factos.

*Armazenamento*: tem que ver com o facto de adicionar novas tabelas de dimensões ou factos ao modelo. Esta operação implica um crescimento significativo do DW, podendo comprometer o seu bom funcionamento em termos de capacidade de processamento ou de resposta às consultas. Conforme é referido por Inmon (1997), a boa capacidade de resposta do DW está relacionada, entre outros factores, com o tamanho e número de tabelas. Para uma boa QD é importante ter em atenção o número de tabelas a usar, assim como o armazenamento de dados ser o menos redundante possível. É ainda necessário perceber a forma como os dados são armazenados nas tabelas em termos de uso de particionamento destas. A arquitectura de armazenamento deve permitir aumentar a capacidade de armazenamento de dados sem degradar a performance.

*Agilidade*: esta dimensão da QD tem que ver com a rapidez de selecção de dados pelos processos de cálculo no DW (boa capacidade de processamento). Para uma boa resposta devem-se usar índices nas tabelas de forma a tornar mais rápido o processamento dos dados, permitindo ao utilizador final aceder aos dados processados o mais rápido possível, tornando os dados *oportunos*. Podem ser ainda englobadas nesta dimensão funções de group-by, union, sum, max, min. O uso destas funções também contribui para a rapidez de resposta do DW, ou seja, para a rapidez com que determinado dado ou conjunto de dados são disponibilizados ao utilizador final.

*Disponibilidade*: de acordo com a dimensão disponibilidade, os dados devem estar disponíveis sempre que o decisor deles necessite. Caso contrário, tudo o resto é posto em causa.

#### 4.3.1.4. Proposta de Dimensões de QD disponibilizados pelas BI

As ferramentas de BI têm a função de aceder aos dados do DW e apresentá-los ao decisor. Não obstante esta realidade, devem garantir total consonância entre a QD no DW e os dados que disponibilizam. Neste contexto, propõe-se de seguida um conjunto de dimensões de qualidade dos dados que estas ferramentas devem seguir.

*Tempo de carregamento de páginas* - O carregamento das páginas pode ter que ver com o excesso de conteúdo nas páginas, com a optimização do código gerado ou com a optimização das imagens usadas. O módulo web de uma ferramenta de BI deve assegurar alguma rapidez de navegação pelas várias páginas e isso consegue-se se o tempo de carregamento das mesmas for mínimo (poucos segundos).

*Compatibilidade com os principais browsers do mercado* - O módulo web de uma ferramenta de BI deve ser compatível com os principais browsers do mercado, isto é, não deve ficar comprometida a disponibilização de dados pelo facto de não se usar um browser corrente no mercado. O uso de etiquetas não conhecidas, na definição de instruções HTML, é um dos problemas que pode ocorrer.

*Estado dos endereços* - Muitas vezes, ao navegar em páginas em ambiente web surgem erros de página não encontrada. Esta é uma característica de falta de qualidade. As ferramentas de BI devem assegurar uma navegação o mais coerente possível e devem preocupar-se em disponibilizar os dados ao decisor permitindo que este "*navegue*" por eles. Como a navegação é feita recorrendo a links, é necessário que estes não estejam quebrados para que se assegure uma ligação activa entre as páginas.

*Erros de programação em HTML* - As ferramentas de BI geram o código de desenho do layout das páginas de forma automática, normalmente em linguagem HTML. É necessário ter em atenção a qualidade do código gerado, nomeadamente quanto a etiquetas em falta ou mal colocadas, tamanho e tipo de fontes ou atributos mal definidos.

*Erros de programação em SQL* - As ferramentas de BI também geram o código de acesso aos dados de forma automática (interrogações). Um problema típico de falta de qualidade destas interrogações é a frequente falta de optimização no acesso aos dados. As interrogações devem usar os índices das tabelas para melhor desempenho e, em caso de não existir índice, devem sugerir a sua criação.

*Disponibilização de dados (Interface)* - As ferramentas de BI devem assegurar um conjunto de princípios de forma a tornar o mais fácil possível a sua utilização pelo decisor. Nielsen, referido por Palma-dos-Reis (1999), em 1993 propôs alguns princípios que, de forma geral, podem ser aplicados no interface das ferramentas de BI. Desses princípios destacam-se o uso de termos conhecidos pelo utilizador, a disponibilização de comandos rápidos (*shortcuts*), o uso de mensagens de erros perceptíveis, a disponibilização de ajuda on-line e o uso de critérios adequados para a utilização de cores e fontes. Acrescenta-se ainda: a forma de disposição dos dados, a facilidade de manipulação dos mesmos e a documentação.

*Flexibilidade* - poder escolher o formato de apresentação que se pretende para visualizar os dados e que pode ir desde tabelas, gráficos ou matrizes. As ferramentas de BI devem

permitir ainda formatar células, construir e executar queries ad-hoc bem como permitir a exportação de dados para outro formato como o Word, Excel ou pdf.

*Documentação dos dados* - Os dados disponibilizados pelas ferramentas de BI devem ser acompanhados pelos metadados mais relevantes nesse contexto. Tais metadados terão de existir em alguma parte do écran, da folha de papel ou em qualquer outro suporte. Exemplos destes metadados podem ser: a data em que o documento foi gerado, os sistemas fonte dos diversos dados elementares, a data de publicação, a data de revisão e a identificação dos autores.

*Capacidade de análise avançada* - As ferramentas de BI têm de ter a capacidade de efectuar operações como a média, a média ponderada, acumulados e ordenação, entre outras.

*Rapidez de cálculo* - as consultas efectuadas através destas ferramentas não podem levar horas ou mesmo vários minutos a apresentar o resultado. É importante que o resultado seja rápido e isso consegue-se quanto melhor for o desenho do DW, mas também quanto maior for a capacidade de processamento do sistema onde o pedido está a ser processado.

#### 4.3.1.5. Metodologias para melhorar a QD no Data Warehouse

Total Data Quality Management (TDQM)

Richard Wang (1998) propõe uma metodologia para melhorar a QD a que dá o nome Total Data Quality Management (TDQM). Esta metodologia assenta, essencialmente, na visão segundo a qual a informação é um produto e os dados a matéria-prima. Esta

metodologia é inspirada no ciclo de melhoria contínua de Deming (*plan, do, check e act*). A metodologia TDQM assenta em 4 fases:

*1 - Definição do ambiente - entender a informação como um produto*. Nesta fase é dada especial atenção ao percurso dos dados na migração para o DW. É elaborado um esquema de produção de dados representando as transformações que os dados vão tendo no processo de migração até chegarem ao DW.

*2 - Medição e avaliação da qualidade dos dados e da informação* através de métricas, dimensões ou indicadores de qualidade integrados no modelo físico.

*3 - Análise das causas da falta de qualidade dos dados e da informação* com recurso a software específico que analise os resultados do ponto 2.

*4 - Desenvolvimento e aperfeiçoamento de acções de melhoria da qualidade dos dados e da informação* como alinhar os fluxos de informação da organização com o sistema de fabricação do produto informação, referido na primeira etapa ou, como refere Sousa-Mendes (2001a), alinhar a informação com as necessidades do negócio.

Esta metodologia pode ser demasiado arriscada, pois na prática existem diferenças significativas entre informação e produto, o que pode comprometer todo o processo. Contudo, se for tido em conta o princípio de que existem diferenças e que estas são consideradas no momento de actuar, esta metodologia pode melhorar a QD no DW.

Total Quality data Management (TQdM)

Larry English (1999) propõe outra metodologia, intitulada Total Quality data Management (TQdM). O autor alerta para o facto desta metodologia não ser um programa ou simplesmente um processo de averiguação ou limpeza de dados. Esta

metodologia surge do facto de todos na organização estarem interdependentes dos dados e da informação e do benefício que esta pode trazer para a organização. Esta metodologia tem como base seis processos chave:

*P6 – Estabelecer um ambiente propício à qualidade dos dados e da informação:*

Este é um processo fundamental pois implica uma mudança de cultura organizacional. É de salientar dois paradigmas, o primeiro, prende-se com o facto da informação ser vista com um produto e os dados como a matéria-prima necessária para a gerar. O segundo, diz respeito ao facto de se aceitarem os custos com a falta de QD como algo normal, isto é, "custos com o negócio". O ambiente e a cultura organizacional pode fomentar e incentivar a falta de qualidade, questões como as barreiras culturais, a pouca percepção em perceber o peso que tem a falta de QD, a pouca percepção em perceber que, melhorando a QD, reduzem-se os custos de negócio e aumenta-se o grau de satisfação dos consumidores, podem contribuir para a faltas de qualidade.

*P1 – Avaliar a definição dos dados e a qualidade da arquitectura de dados:*

Neste processo devem ser tidos em conta um conjunto de procedimentos como: identificar "métricas" de QD, eleger um grupo de dados a avaliar, identificar os intervenientes por categorias de dados, avaliar técnicas de definição de QD, avaliar a arquitectura de dados e qualidade de desenho da base de dados e avaliar o grau de satisfação dos consumidores com a qualidade de definição dos dados.

*P2 – Avaliar a qualidade dos dados e da informação:*

Este processo deve identificar os dados a melhorar, os objectivos e medidas para a QD, o valor dos dados e a cadeia de custos, os ficheiros e os processos a avaliar e os programas fonte para validar os dados. Neste processo devem ainda ser extraídas

amostras para exemplos (estatísticos), reduzindo assim o custo do estudo, e medir a QD resultante interpretando e documentando os resultados obtidos.

*P3 – Medir e avaliar os custos correspondentes à falta de qualidade:*

Este processo inclui um conjunto de procedimentos, são eles, identificar medidas que melhorem o negócio (como por exemplo o que pode aumentar o grau de satisfação do consumidor e o que se pode fazer para reduzir custos); calcular os custos com os dados (no desenvolvimento de infra-estruturas, de valor acrescentado, de redundâncias como BD e aplicações); calcular custos devido a falta de QD e agrupar os custos por categorias. Este processo deve ainda identificar segmentos de consumidores e estabilizar o valor deles, calcular o valor da informação obtida da interpretação dos dados e estabilizar o valor da QD e da informação para medir o valor das oportunidades perdidas. Isto permite perceber não só o estado da QD, mas também o custo actual com a falta de qualidade da informação. Este processo contribui ainda para estimar qual o investimento a fazer nos dois próximos passos, limpeza dos dados e melhoria dos processos de obtenção de informação.

*P4 – Reengenharia e limpeza de dados:*

Neste processo devem ser tidas em conta as tarefas de identificar a origem dos dados que se pretendem para limpeza e reengenharia, extrair e analisar os dados para identificar anomalias e normalizar e corrigir dados em falta (correcção automática ou manual). Após estas tarefas, os dados devem ser comparados e consolidados assim como extraídos das várias fases para se perceber se existem padrões de erros que possam ser usados no processo de melhoramento. De seguida, os dados devem ser mapeados numa estrutura de dados a designar. Por fim, deve assegurar-se que os

processos de auditoria e controlo estão preparados para extrair, transformar, sumariar e carregar os dados das bases de dados operacionais para o DW.

*P5 – Melhorar a qualidade dos processos de obtenção de dados:*

Neste processo é definido o problema que se propõe resolver e, seleccionado, o processo onde se quer melhorar a QD. Para isso, deve desenvolver-se um plano para melhorar a qualidade (analisar causas prováveis) e implementar as alterações correctivas, terminando com os testes. Depois verificar-se o impacto destas medidas de melhoria e unificar as estandardizações sujeitas a alterações.

Esta é uma metodologia bastante abrangente pois parte do princípio de que os trabalhadores estão bastante familiarizados com estas questões, uma vez que é sugerido o conhecimento deles no desempenho das tarefas da metodologia. Pensamos que, para se ter sucesso com a aplicação desta metodologia, além de muito trabalho, é também necessário que os trabalhadores tenham bastante entendimento do processo e estejam comprometidos e motivados. Devem ser possuidores de um nível elevado de conhecimento (trabalhadores do conhecimento). É uma metodologia apropriada para a análise e melhoria da qualidade dos dados nos DW e contribuir desta forma para melhorar a qualidade da informação.

## 4.4. Linhas orientadoras para a migração e limpeza dos dados

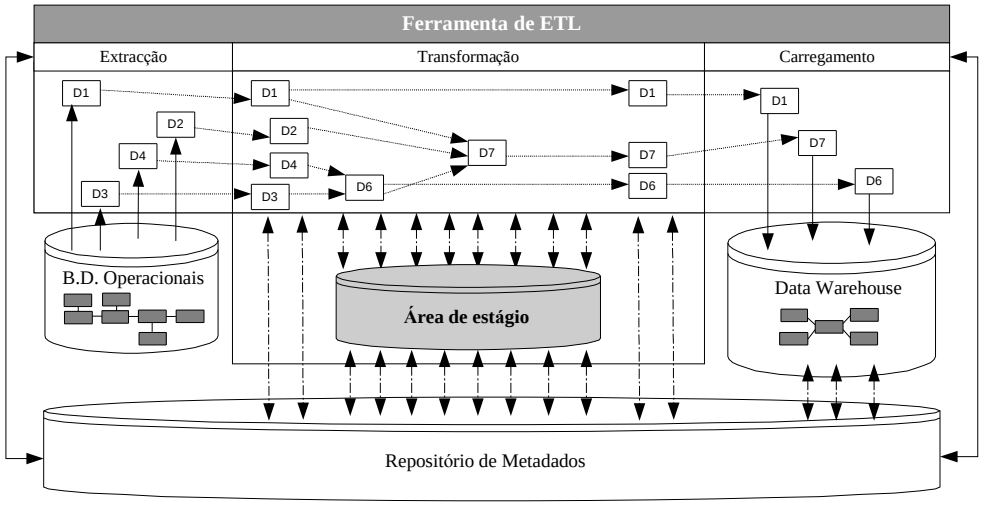
É objectivo, nesta fase, propor um conjunto de linhas orientadoras para a migração dos dados das Bases de Dados Operacionais para o DW. Pretende-se ainda ilustrar o processo de limpeza de dados, tarefa que ocorre na fase de transformação dos dados numa área de ETL, designada por área de estágio (*staging area*).

Soumendra Mohanty (2004), no seu artigo Data Migration Strategies, de Maio, publicado na revista *Data Management Review*, aponta algumas fases que um processo de migração de dados deve respeitar. Com base no seu ponto de vista, juntamente com os de Inmon (1997), Kimball e Caserta (2004) e Moss e Atre (2003) sugere-se no *anexo1* um conjunto de passos e tarefas a ter em consideração nas fases de ***Análise*** dos dados nos sistemas fonte (data source) e no ***Desenho***, ***Teste*** e ***Implementação*** dos processos de ETL. Uma das tarefas das ferramentas de ETL é a *limpeza dos dados*. Esta tarefa ocorre na operação de transformação, onde os dados são sujeitos a processos de limpeza, normalização, cálculo, integração, derivação e agregação.

A *Limpeza de Dados* pode passar por corrigir ou remover dados, marcar os campos com valores nulos, identificar ou remover *outliers* ou ainda resolver inconsistências (Moss e Atre, 2003). De acordo com Berson (1997), os campos com valores nulos podem advir do mau funcionamento do equipamento, do facto de nem sequer terem sido introduzidos, da inconsistência com outros dados registados, de certos dados não serem considerados importantes ou de enganos na entrada de dados. A forma de resolver estas situações passa por preencher manualmente os valores ausentes ou usar uma constante global para representar o valor ausente (Exemplo: "*desconhecido*", "*NULO*"). De acordo com a SAS Institute (2004), a resolução pode passar ainda por usar a média ou a média por classe ou ainda usar o valor mais provável baseado por inferência (fórmula bayesiana ou árvore de decisão). Segundo Kimball e Caserta (2004), os erros com os dados podem advir de problemas com os instrumentos de recolha de dados, problemas de transmissão de dados, limitações tecnológicas ou do não cumprimento de padrões. Para Hall (1999), garantir a integridade e limpeza dos dados é a tarefa mais exigente neste processo. A figura 11 ilustra o fluxo de dados desde as BD operacionais até ao

DW, passando pelos vários processos sofridos na transformação. Todas estas operações têm o apoio do RM.

**Figura 11 - Fluxo de dados no processo de ETL**




Cabe ainda, neste trabalho, exemplificar como se processa a limpeza dos dados. Estes são analisados de forma a eliminar erros ou inconsistências para que cheguem ao DW com a qualidade necessária (Calvanese *et al.*, 1997). Conforme é referido por Galhardas *et al.* (2000), quando se integram dados externos no DW estes muitas vezes trazem inconsistências e, no processo de *Transformação*, é possível criar normalizações nos dados de forma a eliminá-las. A migração de dados pode ter um ou mais sistemas fonte e os possíveis problemas de QD podem ocorrer ao nível da estrutura ou da instância de dados ou ainda nos dois (Rahm e Do Hai, 2000).

Uma fonte de dados

Os dados extraídos de um sistema fonte podem originar vários problemas no momento do carregamento no DW. Ao nível da estrutura de dados, podem ocorrer os seguintes problemas: valores inválidos, dependências não respeitadas, chaves ou integridade referencial violadas (Rahm e Do Hai, 2000). Na tabela 7 exemplificam-se tais situações.

**Tabela 7 - Problemas que podem ocorrer ao nível da estrutura de dados**

| Problema | | Dados com erros | Razão |
|---|---|---|---|
| Atributo | Dados inválidos | Data nascimento = 30.13.1970 | Valor fora do domínio, não existe o mês 13 |
| Registo | Dependência do atributo não é respeitada | Idade = 35<br>Data nascimento = 14.04.1968 | Idade =<br>(data actual – data nascimento)<br>(2004-1968)= 36 |
| Tipo de registo | Chave única violada | Funcionário1<br>Nome = João NSS = 222333444<br>Funcionário2<br>Nome = José NSS = 222333444 | O NSS (número da segurança social) não pode ser o mesmo para diferentes funcionários |
| Origem | Violação da integridade referencial | Funcionário3<br>Nome = Joaquim<br>NSS = 333444555 Cod_dep = 223 | Este código de departamento não está criado na tabela de departamentos |

Ao nível da instância, podem ocorrer erros como: atributos mal definidos; palavras mal escritas; registos duplicados ou referências erradas, conforme observado na tabela 8.

**Tabela 8 - Problemas que podem ocorrer ao nível da instância de dados**

| Problema | | Dados com erros | Razão |
|---|---|---|---|
| Atributo | Dados omissos | Telefone = 999999999 | Foi preenchido o campo sem qualquer critério |
| | Dados mal escritos | Rua = Asinhagga da Cidade | É difícil de controlar para certo tipo de campos |
| | Abreviações sem sentido | Habilitações = LIDG<br>Profissão = TDSI | Não se percebe o que é. |
| | Muitos dados juntos | Nome = António 36 Lisboa CTT | Múltiplos dados no mesmo atributo |
| | Dados incorrectos | Distrito = Alcanena | Não é um distrito |
| Registo | Dependência do atributo não é respeitada | Localidade = Lisboa<br>Código postal = 2380 | O código postal 2380 é de Alcanena e não de Lisboa |
| Tipo de registo | Transposição de palavras | Nome1= J. Andrade<br>Nome2 = João M. | A abreviatura deve ser uniforme e não quando dá mais jeito |
| | Registos duplicados | Funcionário1<br>Nome = José Carlos Matias<br>Funcionário2<br>Nome = J. Carlos Matias | O mesmo funcionário foi introduzido 2 vezes |
| | Registos contraditórios | Funcionário1<br>Nome = Joaquim Almeida<br>Data Nascimento = 21.12.87<br>Funcionário2<br>Nome = Joaquim Almeida<br>Data Nascimento = 21.11.87 | O mesmo funcionário tem duas datas de nascimento diferentes |
| Origem | Referências erradas | FuncionárioX<br>Nome = Joaquim<br>NSS = 333444555<br>Cod_dep = 125 | O departamento 125 existe, mas o funcionário não faz parte desse departamento |

Múltiplas fontes de dados

Os problemas descritos podem aumentar bastante caso existam múltiplas fontes de dados onde cada fonte tem as suas regras, quer ao nível do modelo, quer ao nível da instância de dados (Rahm e Do Hai, 2000). Conforme é referido por Galhardas *et al*. (2000), a limpeza de dados tem aqui um trabalho acrescido na medida que tem de uniformizar os dados das várias origens.

Suponha-se que se pretende extrair os dados das tabelas de colaboradores de dois sistemas fonte distintos e carregar esses dados no DW. As tabelas 9 e 10 apresentam alguns dados das tabelas de funcionários dos sistemas fonte.

**Tabela 9 - Tabela T_Funcionários (Sistema Fonte 1)**

| **T_Funcionários** | | | | |
|---|---|---|---|---|
| *ID* | *Nome* | *Morada* | *Localidade* | *Sexo* |
| 34 | Joaquim Faria | Rua 5 Outubro | 1890 Lisboa | 1 |
| 57 | Cristina Santos | P. Lond | Lisboa | 0 |

**Tabela 10 - Tabela TAB_EMPREGADOS (Sistema Fonte 2)**

| **TAB_EMPREGADOS** | | | | | | |
|---|---|---|---|---|---|---|
| *IDE* | *Primeiro_Nome* | *Apelido* | *Telefone* | *Sexo* | *Morada* | *Cod_postal* |
| 34 | Luís | Sousa | 215555555 | M | Avenida G. Norton de Matos, 543 | 2400 |
| 340 | Cristina | Santos | 214444444 | F | Praça de Londres, Lote 43, 3E, 1200 Lisboa | |

As tabelas T_Funcionários e TAB_EMPREGADOS fazem parte de diferentes sistemas de informação, e como tal a sua estrutura pode não ser a mesma. No momento de migrar os dados relativos aos colaboradores para o DW vão ocorrer problemas ao nível do modelo e da instância de dados. Nas várias operações das ferramentas de ETL estes e outros problemas são seleccionados, originando no DW a tabela 11.

**Tabela 11 - Tabela Colaboradores - (DW)**

| **Colaboradores** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *NO* | *Nome* | *Apelido* | *Sexo* | *Morada* | *Localidade* | *CPostal* | *Tel.* | *ID* | *IDE* |
| 1 | Joaquim | Faria | M | Rua 5 Outubro | Lisboa | 1890 | *NULO* | 34 | |
| 2 | Luís | Sousa | M | Avenida G. Norton de Matos | Leiria | 2400 | 215555555 | | 34 |
| 3 | Cristina | Santos | F | Praça de Londres, Lote 43, 3E | Lisboa | 1200 | 214444444 | 57 | 340 |

Conforme se constata da observação das tabelas, existiam problemas de vária ordem, nomeadamente o campo sexo estava definido numa tabela com o domínio (M,F) e noutra com o domínio (1,0). Os nomes de alguns campos, que supostamente guardam o mesmo dado, têm designações diferentes e alguns dados estão em campos menos apropriados. No caso dos colaboradores, um deles está registado nas duas tabelas e o mesmo número sequencial é atribuído a diferentes colaboradores.

Com a migração dos dados para o DW estas situações ficam corrigidas levando ao melhoramento da QD. Basta constatar que antes o número de colaboradores não estava correcto porque um deles estava nos dois sistemas, as moradas não estavam completas nem os códigos postais estavam sempre inseridos no local correcto. As colunas ID e IDE foram mantidas para que, caso seja necessário, se consiga relacionar estes registos com as suas fontes. Este é um pequeno exemplo do que se pode fazer para aumentar a qualidade dos dados nas organizações.

# 5. Arquitectura para um Ambiente Analítico

Nos anos 90, muitas foram as organizações que adquiriram novos sistemas de informação, nomeadamente ERPs ou CRMs, com o intuito de, por um lado, acabar com os sistemas legados ou resolver o problema do ***bug*** do ano 2000 ou mesmo a adopção à moeda única europeia (EURO) e, por outro, melhorar a QD dos seus sistemas. A verdade é que, após estas tarefas concluídas, continua a verificar-se uma grande necessidade de melhoria da QD, e por conseguinte da qualidade de informação.

Este problema pode ter que ver com o facto de não se ter usado uma arquitectura capaz de suportar uma ferramenta de ETL que assegurasse a migração dos dados dos vários sistemas, de forma a garantir a qualidade dos dados no DW (Carreira e Galhardas, 2004). Os dados foram migrados, mas os problemas que existiam anteriormente relativos à falta de qualidade não só se mantiveram como, nalguns casos, ainda se agravaram, uma vez que estes ficaram "***misturados***" não se garantindo a sua total correcção ou documentação.

Neste sentido, verifica-se a necessidade de uma arquitectura capaz de lidar com tal problema de forma a garantir a migração dos dados dos vários sistemas fonte para um sistema centralizado capaz de os disponibilizar às ferramentas de BI. Conforme sugerido no capítulo 3, tal sistema deve ser o DW, e no processo de migração deve ter-se em conta a documentação dos dados e dos processos no repositório de metadados.

Existem ainda outras ***necessidades***, que agem em conjunto para motivar as organizações à adopção de uma ***Arquitectura para um Ambiente Analítico*** que permita às ferramentas

de BI desempenhar o seu papel. As necessidades que motivam a adopção deste ambiente são:

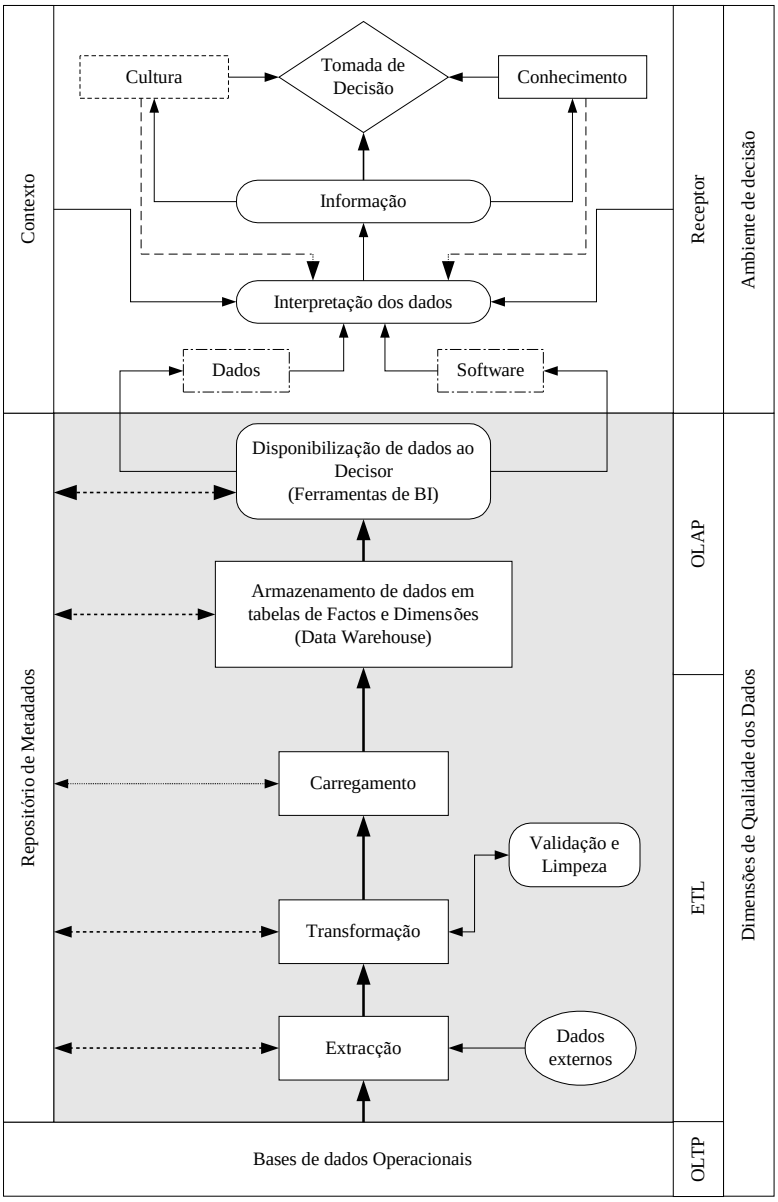
- Necessidade de centralização dos dados das aplicações produtivas distribuídas pela organização, independentemente das diferentes tecnologias usadas.
- Necessidade de usar um sistema em que possa confiar a centralização dos seus dados.
- Necessidade da organização garantir a qualidade dos dados no processo de centralização dos mesmos.
- Necessidade de ferramentas capazes de processar os dados.
- Necessidade de disponibilizar os dados aos gestores e aos analistas de negócio.

Consideradas estas questões, as ferramentas de BI estão em condições de usar os dados armazenados no DW. Estas ferramentas, como já referido, proporcionam uma melhor capacidade de visão dos mercados e das operações internas das organizações, permitindo que estas reajam rapidamente às mudanças no ambiente e se preparem para o futuro. Mas, para que desempenhem o seu real papel, é necessário que as organizações lhes garantam a qualidade dos dados. Neste contexto, a adopção de uma *Arquitectura para um Ambiente Analítico* é a forma mais segura e rápida de obter esta qualidade, e assim poder obter vantagens competitivas. O uso de ferramentas de BI tornou-se uma necessidade competitiva, mas os dados que usam devem apresentar qualidade, só assim os gestores poderão tomar decisões baseadas em informação de qualidade. Contudo, deve ter-se em conta que estas decisões não dependem única e exclusivamente dos dados, uma vez que, conforme já referido no capítulo 2, existem outros factores que influenciam a tomada de decisão.

## 5.1. A Framework para um Ambiente Analítico

Esta framework é o ponto de partida para a construção de uma Arquitectura para um Ambiente Analítico e pretende-se materializar e arrumar tecnologicamente os vários conceitos descritos ao longo deste trabalho. Espelha o fluxo dos dados e processos a que estes estão sujeitos, desde os sistemas fonte até ao seu uso na tomada de decisão.

**Figura 12 - Framework para um Ambiente Analítico**



## 5.2. A Arquitectura para um Ambiente Analítico

A *Arquitectura para um Ambiente Analítico* engloba os sistemas fonte (normalmente dados operacionais e dados externos relevantes para a gestão), as ferramentas de extracção, transformação e carregamento de dados (ETL), o repositório de dados (DW), o RM e, por último, as ferramentas de BI que disponibilizam os dados aos utilizadores para que estes, no processo de decisão, possam obter informação de qualidade.

Considera-se existirem três níveis na *Arquitectura para um Ambiente Analítico*. O primeiro, acolhe as aplicações operacionais da organização com os respectivos dados necessários ao funcionamento do dia-a-dia, sendo neste estágio que se encontram os dados a tratar pelas ferramentas de ETL com o objectivo dos carregar no DW.

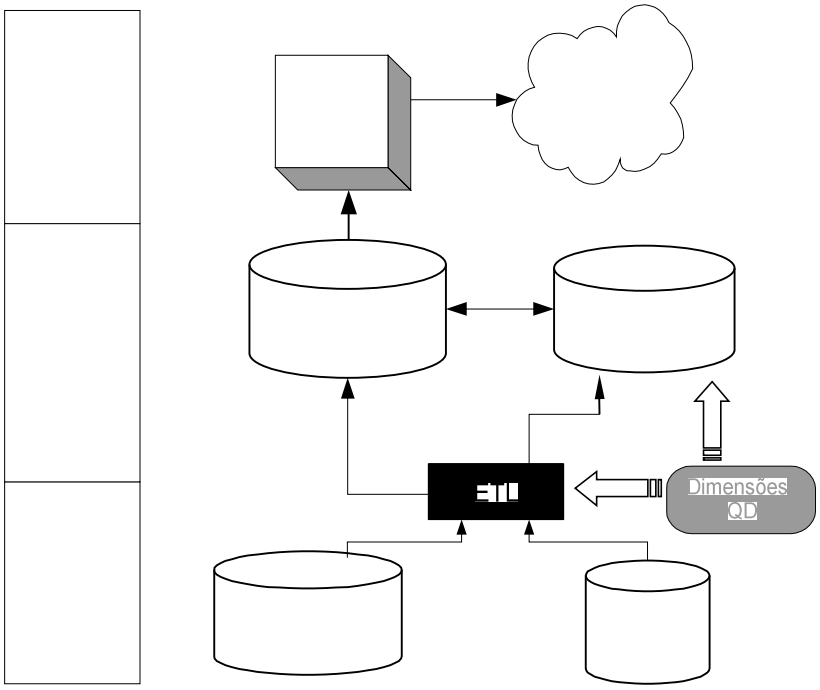
No segundo estágio, encontra-se o processo de centralização dos dados que, como já referido, recorre ao uso de ferramentas de ETL. Tais ferramentas fazem a "*ponte*" entre estes dois níveis, já que extraem dados de vários Sistemas Operacionais transformando-os e inserindo-os no DW. Para além desta operação, insere ou actualiza, no repositório de metadados, todas as operações a que os dados estiveram sujeitos no processo de centralização. As ferramentas de ETL usadas devem ter em atenção a garantia da qualidade dos dados e para isso devem permitir a aplicação de dimensões de qualidade[4], ou seja, devem disponibilizar um ambiente onde se possa analisar e melhorar a QD de acordo com as várias dimensões. O RM servirá, como já referido, para documentar os

[4] Essas dimensões foram as propostas no capítulo 4 para o Data Warehouse.

dados e processos associados, as regras de negócio, os processos de carregamento e a periodicidade de refrescamento dos dados das BD operacionais para o DW.

Por fim, e no terceiro estágio, situam-se as ferramentas de BI, alimentadas por dados fornecidos pelo DW. Aqui, as ferramentas de BI processam e disponibilizam os dados ao decisor de acordo com os seus pedidos, que podem ser consultas (QUERY), relatórios (REPORT) ou análises multidimensionais (OLAP). Na figura 13 ilustra-se a *Arquitectura para um Ambiente* Analítico, espelhando os três níveis.

**Figura 13 - Arquitectura para um Ambiente Analítico**

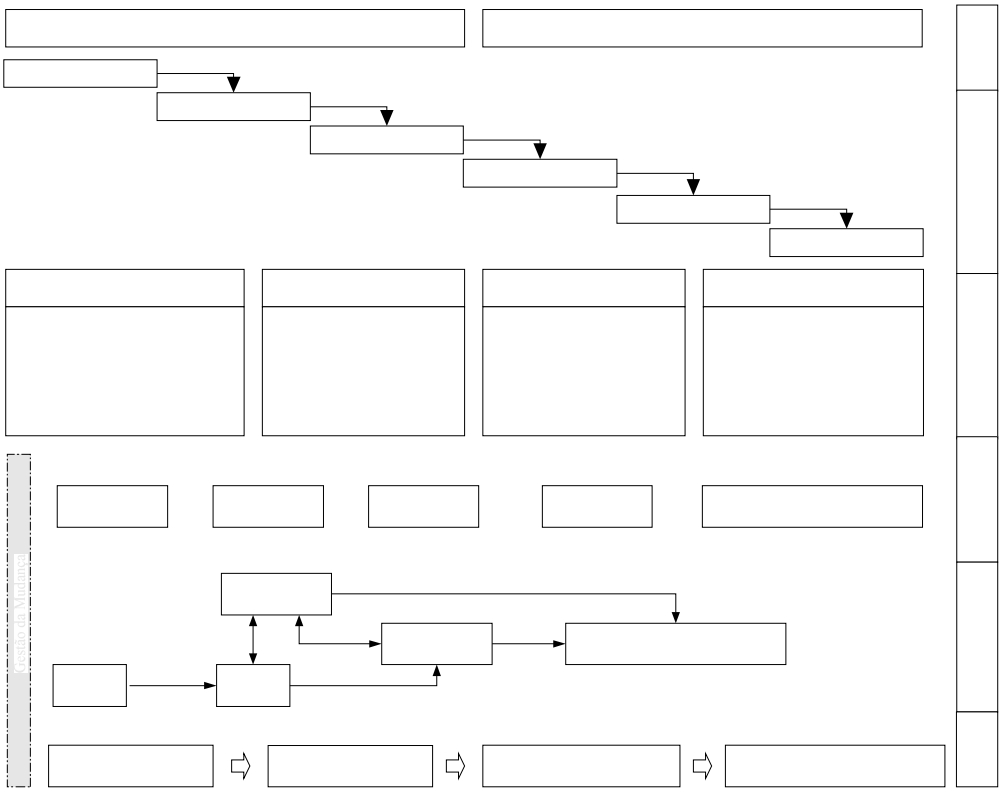



O sucesso das ferramentas de BI está fortemente relacionado com a qualidade dos dados que usam e, neste sentido, a solução da *Arquitectura para um Ambiente Analítico* tem em vista a garantia dessa qualidade. Esta arquitectura permite um melhor tratamento dos dados dos sistemas fonte para o DW, já que pelo meio tem o apoio das ferramentas de ETL assim como a ajuda do RM para documentar todas as operações.

## 5.3. Plano de construção de uma Arquitectura para um Ambiente Analítico

Não é objectivo deste trabalho detalhar a construção de um *Ambiente Analítico* contudo, a figura 14 pretende dar um contributo para uma possível abordagem.

**Figura 14 - Plano de Construção de uma Arquitectura para um Ambiente Analítico**




Inicia-se com a necessidade da organização disponibilizar dados de qualidade (*Justificação e Avaliação Ex-Ante*), daí planear e apresentar uma solução que responda a tal necessidade (*Plano*). Após o plano efectuado deve ser prioridade da organização fazer um levantamento a aspectos como: aspectos de gestão, aspectos tecnológicos, aspectos humanos e aspectos da cultura organizacional (*Análise do Negócio*). Findo esta fase, inicia-se o *Desenho e Desenvolvimento* da solução e respectiva *Implementação* da Arquitectura estando, nestas fases, presente a Gestão da Mudança. Por fim, deve ser feita uma avaliação dos resultados obtidos (*Avaliação Ex-Post*).

## 5.4. Por que pode falhar a Implementação da Arquitectura

Pode-se desenhar, construir e implementar uma *Arquitectura para um Ambiente Analítico* e usar boas ferramentas de BI contudo, muitas vezes estes projectos falham. Segundo Larissa T. Moss e Shaku Atre (2003), as organizações, ao enveredarem pela implementação de ferramentas de BI, devem ter em conta um conjunto de desafios críticos a que é necessário atender. Estes autores citam 10 desses desafios:

- Capacidade de reconhecer que este tipo de projecto engloba toda a organização.
- Necessidade fulcral de centralização dos dados.
- Ter um Sponsor ao mais alto nível.
- Ter os responsáveis pelos negócios motivados e disponíveis.
- Alto grau de qualificação do pessoal.
- Conhecimentos sólidos no desenvolvimento de software.
- Utilização de metodologias bem testadas.
- Boa Análise de negócio.
- Boa avaliação de impacto dos dados no sistema.
- Perceber bem a necessidade e importância dos metadados.
- Não ter excesso de confiança nos métodos e nas ferramentas (síndroma da *Bala de Prata*[5]).

Estas preocupações dizem respeito não só à implementação de ferramentas de BI, como também à construção de um ambiente propício onde estas possam usar todas as suas potencialidades. Neste sentido, a *Arquitectura para um Ambiente Analítico* pode

[5] Frederick P. Brooks autor, em 1987, da expressão "*There is no silver bullet*".

contribuir para o sucesso da organização, já que o decisor vai tomar as decisões com base em informação obtida através dos dados disponibilizados pelas ferramentas de BI. Estas, por sua vez, vão usar os dados vindos das bases de dados operacionais e carregados no DW, com a ajuda de ferramentas de ETL, e todo este processo é documentado no RM.

Para reforçar a importância da QD na implementação da arquitectura salientamos a conclusão de Eckerson (2002) que, após a realização do inquérito efectuado pelo Data Warehouse Institute a 647 organizações de vários países e diversas áreas, afirma que as organizações investem milhões de dólares em ferramentas de BI mas não dedicam a devida atenção à qualidade dos seus dados. Segundo este estudo, apenas 26% das organizações tiveram a qualidade dos dados como preocupação nos últimos 3 a 5 anos. Mais preocupante é o facto de 52% das organizações não terem qualquer plano no que diz respeito à qualidade dos dados. Apenas 12% das organizações utilizam ferramentas para melhorar a qualidade dos dados e 11% planeiam usá-las nos próximos 12 meses.

Podem-se retirar daqui algumas conclusões, nomeadamente, a pouca importância dada à QD por parte da maior parte das organizações. Muitas pensam que é suficiente investir milhões de dólares em tecnologia para lhes resolver os problemas, quando na verdade o sucesso dessas tecnologias depende de factores como os dados (QD), aos quais não é dada a devida importância. Neste sentido, torna-se evidente a necessidade de as organizações investirem nesta área. Robert Brauer (2001), presidente da DataFlux, uma subsidiária do SAS Institute dedicada à "***limpeza dos dados***", afirma que a qualidade

destes e as ferramentas de BI devem andar de mãos dadas, pois o sucesso destas ferramentas depende em muito da QD.

Muitos projectos falham pelo facto das organizações não optarem pelas ferramentas de ETL mais adequadas. A título de exemplo, pode-se apontar o facto de não se ter em conta as compatibilidades entre as ferramentas e os sistemas fonte e destino. A escolha de ferramentas de ETL adequadas é muito importante para o sucesso do projecto. Se tal escolha não for bem feita, o sucesso da *Arquitectura para um Ambiente Analítico* poderá ficar seriamente comprometido. As ferramentas de ETL são o coração de todo o sistema e, conforme é referido por Inmon (1997), consomem entre 65% a 80% do tempo gasto na construção de todo o sistema. Estas ferramentas são o elo de ligação entre o ambiente OLTP e o ambiente OLAP.

### 5.4.1. Ferramentas de ETL e de BI - Desenvolvimento à medida ou aquisição?

No processo de eleição destas ferramentas, ou mesmo das ferramentas de BI, duas situações são possíveis: uma, o desenvolvimento à medida, outra, a aquisição no mercado. Contudo, independentemente da opção a tomar, há que elaborar uma criteriosa análise de custos, benefícios e viabilidade, pois ambas as opções oferecem vantagens e inconvenientes.

Ao optar pelo desenvolvimento à medida, a organização necessita de colaboradores capazes de desenvolver o software, ou então tem que recorrer ao desenvolvimento externo através da contratação de técnicos. É fundamental utilizar uma metodologia de desenvolvimento bem testada e envolver toda a organização através da constituição de equipas multi-disciplinares. Em contrapartida, optar pela aquisição, implica criar um

modelo de gestão do projecto para que se possa testar, validar e implementar o software adquirido.

Defendemos que o caminho mais seguro é a aquisição de boas ferramentas no mercado que garantam uma total compatibilidade entre os sistemas fonte, o DW e o RM. Kimball e Caserta (2004) salientam a vantagem de estas ferramentas estarem já munidas de RM compatíveis com outros sistemas. De facto, estas ferramentas devem disponibilizar o acesso a vários SGBD's como Oracle, DB2, Informix, SQLServer. Devem, igualmente, disponibilizar um bom ambiente gráfico que permita planear e executar o processo de forma amigável e posterior monitorização. Normalmente, as empresas fornecedoras de software disponibilizam uma versão para testes, onde os potenciais utilizadores podem testar e classificar de acordo com os seus critérios. Podem ainda ser marcadas reuniões com os representantes para possíveis esclarecimentos, assim como aceder a documentação vária sobre as ferramentas. No final torna-se mais fácil escolher qual a adquirir, tendo a vantagem de estar já "aprovada" pelos potenciais utilizadores.

Com a aquisição, os custos podem ser mais controlados, contribuindo para reduzir as hipóteses de surpresas desagradáveis no final do projecto (Madsen, 2004). Contudo, há que ter em conta que o facto de adquirir um software por determinado valor não quer dizer que seja esse o valor final a pagar. De facto, verifica-se, com frequência, um acréscimo no valor previsto porque, ao longo do projecto de implementação é detectada a necessidade de mais um módulo, uma funcionalidade ou outra qualquer extensão. A vantagem é que estas extensões já existem, já foram testadas e, têm um custo conhecido. Ao desenvolver o software, essas necessidades também de verificam, embora com a desvantagem de nunca terem sido testadas e não se saber à partida qual o seu custo.

# 6. Aplicação da Arquitectura para um Ambiente Analítico numa organização concreta

Neste capítulo pretende-se analisar um caso concreto, recorrendo para isso à área de Track & Trace dos CTT. Terá de se começar por pensar como melhorar a QD nos CTT, e para tal parte-se da cadeia de valor da empresa no que respeita aos seus dados e aplicações. Embora a cadeia de valor tenha tido sempre uma importância fundamental no estabelecimento das estratégias organizacionais, é sobretudo neste momento em que a concorrência começa a ditar regras por força da *liberalização* do mercado que devem aumentar as preocupações com a QD e com a forma de gestão dos dados. Muitos sistemas, directa ou indirectamente, influenciam a qualidade da resposta ao cliente e, por conseguinte, a satisfação deste.

## 6.1. Levantamento da situação actual

*"(...) o sector postal passou a ser encarado pelo mercado, como um custo passível de racionalização, observando-se na maioria dos Operadores Públicos Postais, uma quebra na sua actividade tradicional. (...) Aceleraram-se os efeitos do processo de concentração empresarial, particularmente nos sectores financeiros e segurador. Este processo tem significado reduções significativas no volume de correio enviado. A concorrência intensificou-se, quer ao nível tecnológico (...), quer ao nível físico com a entrada de novos operadores nos segmentos mais rentáveis do negócio."*

Dr. Carlos Horta e Costa
Relatório e Contas, 2002

A realidade empresarial da indústria postal, espelhada no relatório referido na citação anterior, enfrenta crescentes níveis de competitividade, sendo um deles a capacidade de

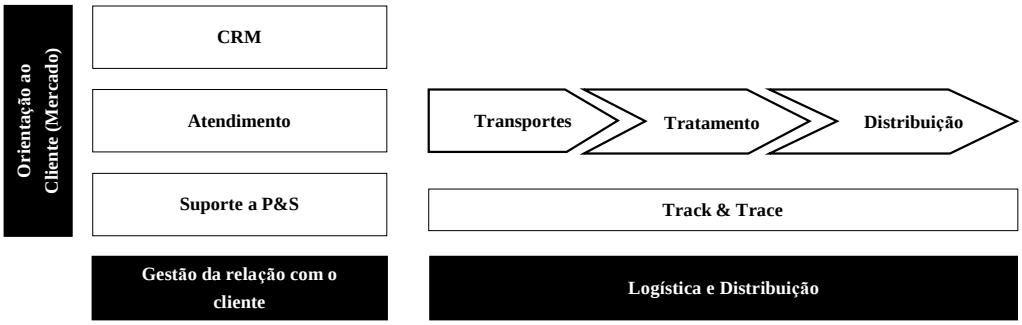
se manter competitivo dentro do seu mercado, sobretudo em épocas de menor vigor macro-económico. Esta competitividade pode ser conseguida através de maiores níveis de eficiência, superior eficácia nos investimentos realizados, melhor percepção da dinâmica do mercado, conduzindo às novas necessidades dos clientes ou ao melhor relacionamento com os fornecedores. E nestes ambientes de incerteza, é necessário desenvolver um *ambiente analítico* concebido especialmente para facilitar a obtenção de informação e uso de conhecimento.

Os sistemas de informação desempenham um papel central e insubstituível em todo este modelo, não bastando que os dados existam e sejam disponibilizados, mas exigindo-se formalmente que, em cada fase da cadeia de valor eles, estejam disponíveis de modo a poderem ser partilhados e processados pelos restantes colaboradores com vista à obtenção de informação para a decisão.

Citando ainda o Dr. Carlos Horta e Costa no mesmo relatório, *"Pretende-se (...) obter uma medida de actividade, numa base diária, o que até agora é impossível"*. Justifica-se assim o desenvolvimento de uma componente de Gestão Estratégica e Corporativa que permita obter indicadores relevantes da actividade dos Correios. A arquitectura apresentada no capítulo anterior pode contribuir para alcançar este e outros objectivos.

O estudo à abordagem dos sistemas de informação nos CTT vai ser efectuado apenas no sub-sistema Track & Trace, devido à dimensão dos Correios e, sempre que possível, numa orientação ao cliente. Apresenta-se na figura 15 a cadeia de valor da empresa que espelha o ambiente produtivo postal dos CTT.

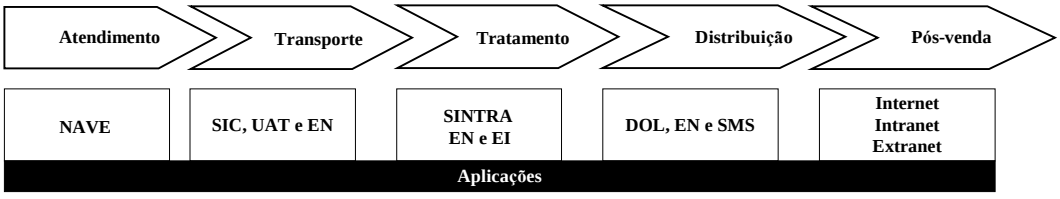
**Figura 15 - O SI orientado ao cliente**




Conforme se pode observar, a orientação ao cliente divide-se em duas vertentes: a primeira, dizendo respeito à relação com o cliente de forma directa, através do *CRM*, do atendimento personalizado e do suporte a produtos e serviços (P&S); a segunda, relativa ao nível produtivo do negócio, ou seja, a logística e a distribuição de objectos postais. Estas duas vertentes ilustram a cadeia de valor dos CTT desde a aceitação de objectos postais e serviços até à sua distribuição e consequente entrega.

O sub-sistema **Track & Trace** é um sector chave dos CTT porque, grosso modo, é o sistema que permite gerir o acompanhamento dos objectos postais desde a sua aceitação (nas estações) até à sua entrega (cliente final). Tem associado um vasto conjunto de aplicações, desenvolvidas em diferentes tecnologias conforme se pode verificar na figura 16.

**Figura 16 - As Aplicações na Cadeia de Valor**

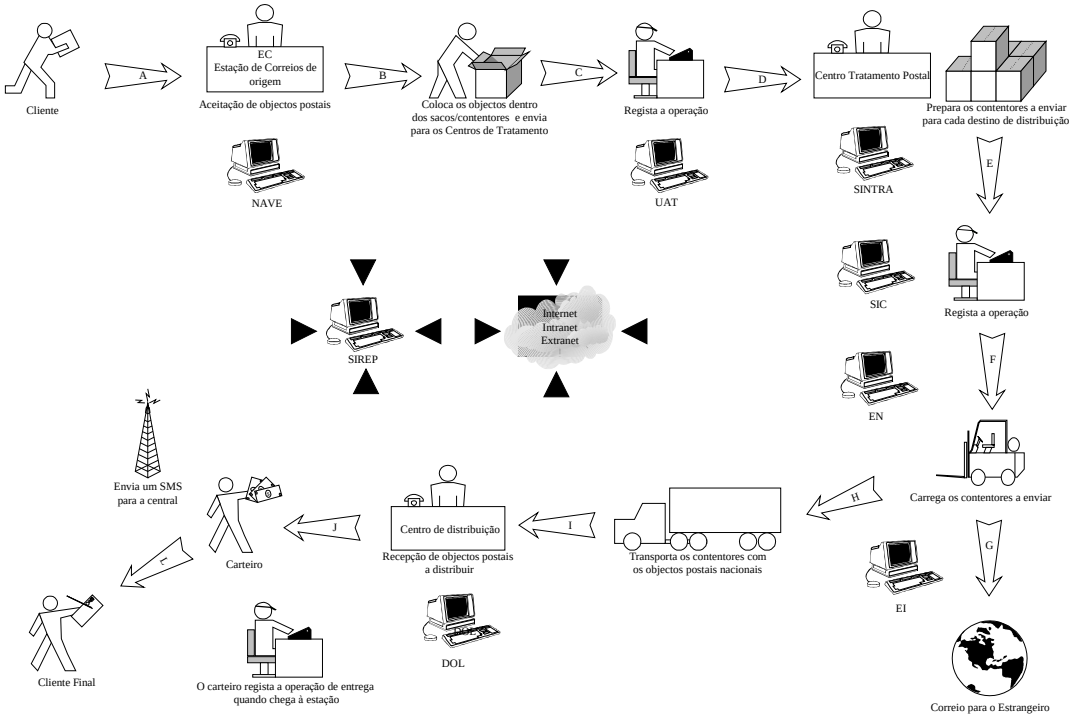


- A aplicação ***NAVE***, desenvolvida à medida, tem a função de gerir o atendimento, encontrando-se portanto instalada nas Estações de Correios. Através dela processa-se o registo de objectos postais, quer a *aceitação* quer a *entrega* ao cliente final.
- A aplicação ***SIC*** tem como função gerir o transporte dos objectos postais: horários, locais, origem e destino de um determinado objecto postal.
- A aplicação ***UAT*** tem como função a gestão de contentores (normalmente os objectos postais são transportados dentro destes contentores).
- A aplicação ***EN*** tem por função o registo de todos os objectos e contentores incluídos em cada meio de transporte (para os vários destinos destino).
- A aplicação ***SINTRA*** tem por função a afectação e optimização de recursos, através de uma correcta planificação do trabalho.
- A aplicação ***EI*** é responsável pelo registo dos objectos postais que se destinam ao estrangeiro.
- A aplicação ***DOL*** apoia a distribuição de objectos ao cliente final (a aplicação ***EN*** também disponibiliza esta função em alguns locais do país).
- A aplicação ***SMS*** tem como função o apoio na distribuição de objectos ao cliente final, mas numa vertente on-line. Genericamente, esta aplicação está inserida num terminal móvel onde o carteiro regista a entrega de um objecto postal e, após efectuado o registo, o terminal envia os dados através de um colector de mensagens SMS para uma Base de Dados.
- A aplicação ***SIREP*** tem a função de centralizar os dados provenientes das aplicações ***EI***, ***EN***, ***DOL*** e ***SMS***, e indirectamente de algumas das outras aplicações.

Algumas destas aplicações estão na Intranet/Extranet da empresa com o objectivo de disponibilizarem dados, sobre os objectos postais, às operações e aos serviços pós-venda. Estes dados podem ser, por exemplo, o dia em que o objecto foi recepcionado na EC, o dia em que foi entregue ao destinatário, por quem e a que horas.

Após explicação da principal funcionalidade de cada aplicação, a figura 17 ilustra o modo como as aplicações são usadas em todo o país, no dia-a-dia dos Correios.

**Figura 17 - O uso das aplicações no dia-a-dia**




O cliente dirige-se a uma Estação dos Correios (EC) e entrega os objectos postais. O funcionário faz a respectiva aceitação e procede ao registo da operação no sistema NAVE. Nesta operação, o cliente paga a importância correspondente ao serviço. O sistema informático regista esta operação com um evento, neste caso o "*evento A*", que corresponde ao primeiro evento postal do objecto.

Após esta operação a EC procede à colocação do objecto num contentor com a finalidade de o enviar para o Centro de Tratamento (CT). Nesta operação é registado o "*evento B*" na aplicação UAT. Quando o objecto é enviado para o CT é gerado o "*evento C*" na mesma aplicação.

No Centro de Tratamento, antes de chegar o objecto postal fisicamente, chegam os dados relativos a esse objecto, via FTP, com a finalidade de a aplicação SINTRA proceder à planificação de trabalho e alocação de recursos de acordo com a quantidade de objectos, sendo neste processo criado o "*evento D*". Os objectos postais vão chegando ao longo da noite e vão sendo separados de acordo com o seu destino, e gera-se o "*evento E*" na aplicação SIC. Quando esta operação estiver concluída, é registado mais um evento na aplicação EN ou EI, consoante o objecto for para distribuir em Portugal ou no estrangeiro. Caso seja na aplicação EN é gerado o "*evento F*", que corresponde à expedição do objecto para o Centro de Distribuição respectivo. Caso seja a aplicação EI é gerado o "*evento G*", que corresponde à expedição do objecto para um determinado país. Quando o objecto sai efectivamente das instalações, a aplicação SIC regista o "*evento H*", que corresponde ao envio do objecto por um determinado meio de transporte.

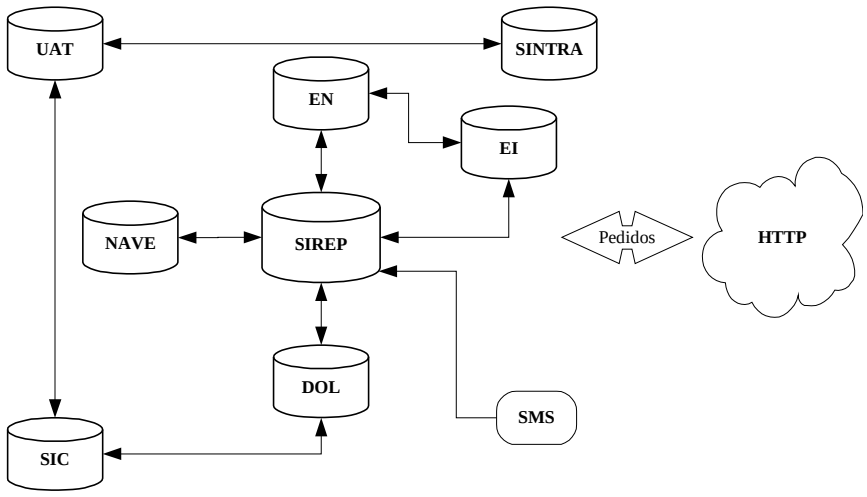
Quando o objecto postal chega ao Centro de Distribuição, e após ser aí recepcionado, é gerado na aplicação DOL o "*evento I*". Segue-se a separação e entrega do objecto a um carteiro que o distribui e é registado na mesma aplicação o "*evento J*", que corresponde à saída do objecto para entrega ao cliente final.

Este processo finaliza com a entrega do objecto ao cliente, que corresponde ao "*evento L*", evento esse que pode ser gerado de duas formas: se o carteiro tiver um terminal móvel (SPT) envia esse evento por SMS para a aplicação de SMS; se não tiver terminal móvel, o evento é registado na aplicação DOL no regresso do Carteiro ao Centro de Distribuição.

A finalidade de registar estes eventos postais é permitir a qualquer momento fazer o *trace* do objecto portal de forma a visualizar o seu percurso.

Quanto à arquitectura de bases de dados resultante destas aplicações pode dizer-se que cada aplicação dá origem a um repositório de dados, ou seja, a uma base de dados. Constata-se ainda que muitas destas aplicações estão espalhadas pelo país e cada uma tem o seu repositório associado. O acesso às várias aplicações para efeitos de consulta é via pedidos HTTP. A arquitectura de base de dados final é a espelhada na figura 18.

**Figura 18 - Arquitectura Actual do Track & Trace e fluxo de dados**



## 6.2. Comentários à situação actual

Os CTT não estão alheios ao ambiente complexo em que muitas organizações se inserem actualmente e também eles operam cada vez mais em tal ambiente. Além da globalização de mercados, os CTT enfrentam a liberalização do mercado postal, que os relega definitivamente para um ambiente de decisão complexo.

A arquitectura de sistemas de informação no sub-sistema Track & Trace, analisada neste levantamento, reflecte uma arquitectura característica de ambientes simples. O desenvolvimento de aplicações é feito no sentido de servir as necessidades das operações e a gestão nem sempre usa os dados lá armazenados.

Mas de acordo com tudo o que foi dito ao longo deste trabalho, e confirmando-se a evolução dos CTT para um mercado altamente competitivo e complexo, torna-se necessário alterar a arquitectura de sistemas de informação por forma a responder aos novos desafios.

A qualidade dos dados e a sua centralização é um dos desafios que a empresa enfrenta e, neste sentido, propõe-se um conjunto de tópicos a que é necessário dar atenção:

- Ausência de um repositório central (Data Warehouse).
- Ausência de um repositório de metadados
- Tabelas de referência desactualizadas e descentralizadas.
- Ausência de um integrador aplicacional (está a ser implementado o BizTalk Server da Microsoft).

- Desenvolvimento de sistemas de informação numa perspectiva global e não de forma isolada como até há bem pouco tempo.

A atenção a dar a estes tópicos deverá fomentar uma nova abordagem - a das ferramentas de *Business Intelligence*. Neste sentido, é necessária uma nova arquitectura de sistemas de informação que possa dar resposta em tempo e com qualidade a um conjunto de questões que a actual arquitectura, pela forma como foi sendo desenvolvida e pelas necessidades de então, não permite. Como exemplo podem referir-se algumas questões a que actualmente não se conseguem responder em tempo e com qualidade:

- Produtos mais vendidos.
- Estações mais rentáveis.
- Total facturado diariamente, por produto, por estação, por distrito.
- Melhores clientes.

## 6.3. Proposta de melhoria com a Arquitectura para um Ambiente Analítico

De acordo com o ambiente descrito, e de forma a responder aos presentes desafios e necessidades de negócio, propõe-se uma infra-estrutura que permita:

- Desenvolver uma visão centrada no cliente e nos produtos.
- Administrar diferentes dimensões do negócio.
- Optimizar cada interacção relativa ao negócio.
- Melhorar a capacidade analítica escalar ou georeferênciada.
- Viabilizar processos de segmentação.
- Quantificar o valor do negócio.
- Optimizar o processo de orçamentação, seja em planeamento ou execução.

- Facilitar a gestão na definição de acções, como novos produtos ou estações.

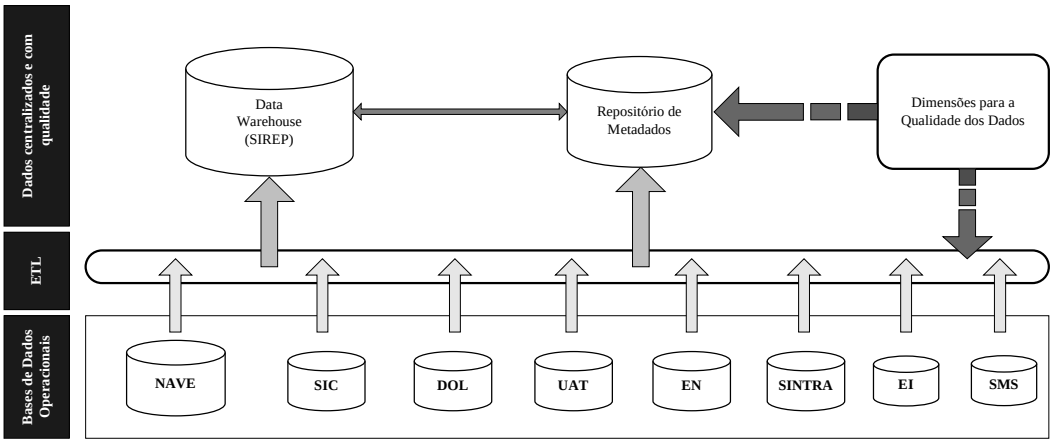
A solução proposta consiste na implementação de um modelo de dados centrado no negócio, principalmente no cliente e nos produtos, e devidamente caracterizado com todos os dados específicos à actividade postal – procura postal, infra-estruturas, relação com terceiros e perfis de serviços. Esta solução está desenhada de forma a permitir o desenvolvimento de processos de exploração de dados no âmbito conceptual de *Business Intelligence – reporting* com ferramentas cliente SQL, a modelação multidimensional para exploração OLAP, a criação, caso se justifique, de *Data Marts* específicos de unidades de negócio, a infra-estrutura nativa de *data mining* e a exploração visual com auxílio de informação georeferenciada, permitindo responder a questões como:

- Conheço os meus clientes? Quais as suas necessidades?
- Qual o seu perfil relativamente a consumo, frequência e pagamento?
- Qual a melhor oferta que poderei realizar a determinada entidade? Qual o plano de tarifário mais adequado?
- Como conseguir aumentar o consumo dos meus clientes?
- Como evitar processos de abandono dos clientes mais rentáveis?
- As iniciativas projectadas na organização estão a trazer valor?
- Será que as campanhas são efectivas? E as promoções, estão a chegar aos clientes desejados?

A nova solução parte da arquitectura actual e assume a aplicação SIREP como sendo o DW, e todos os outros sistemas como aplicações táctico-operacionais. Assim, e

recorrendo ao uso quer de uma ferramenta de ETL, quer de um repositório de metadados, chega-se à proposta da *Arquitectura para um Ambiente Analítico.* Deste trabalho resultam dados tratados, centralizados e documentados, conforme ilustrado na figura 19.

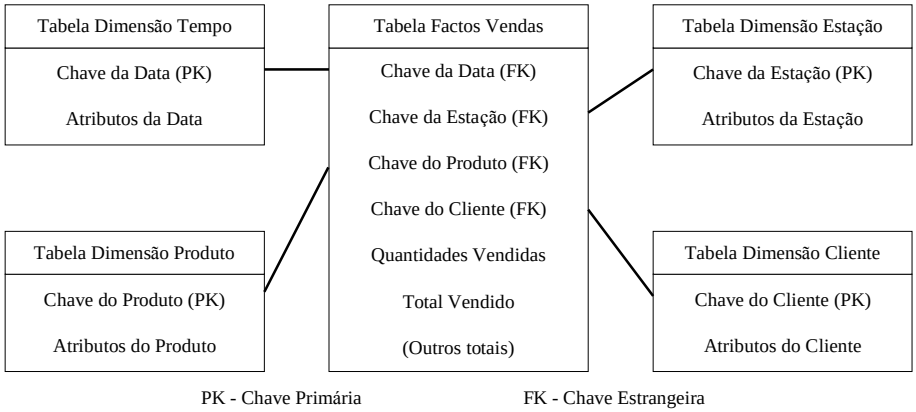
**Figura 19 - Nova Arquitectura para um Ambiente Analítico no Track & Trace**




Para tornar mais fácil de perceber a implementação desta arquitectura, centremo-nos na aplicação **NAVE**. Esta aplicação tem um conjunto de funcionalidades, sendo a principal o registo de objectos postais no momento da sua aceitação nas estações de correios. Regista, entre outras coisas, dados do cliente, do produto vendido, da estação de origem e da estação de destino, isto é, registam-se os dados da venda.

Em termos de DW apresenta-se o modelo necessário para suportar a estrutura descrita. Esta estrutura é composta por quatro tabelas de dimensão (*tempo, produto, estação e cliente*) e uma tabela de factos (*vendas*), como se ilustra na figura 20.

**Figura 20 - Modelo em Estrela: tabelas de factos e dimensões**




Com esta estrutura, as tabelas de dimensão possuem os dados relativos à data de venda de um determinado produto, os dados do cliente, os dados da estação de origem (onde ocorreu a aceitação do produto) e da estação de destino (onde vai ocorrer a entrega ao cliente final). Na tabela de factos encontram-se os dados das vendas, que permitem conhecer o total facturado por dia, por cliente e por estação. Permitem ainda conhecer as vendas por produto (correio registado, correio urgente, correspondências, encomendas) e o número de objectos postais para cada estação de destino. Com estas tabelas "*populadas*" de dados das várias estações de correios do país, e aplicando ferramentas de BI, torna-se possível uma análise do negócio . Podem ainda existir outras tabelas de dimensão como a tabela transporte (da aplicação SIC), que por questões de facilitar a compreensão não estão representadas no modelo acima. Contudo, esta tabela de dimensão permitiria analisar o negócio tendo em atenção o tipo de transporte usado para transportar os objectos postais e igualmente permitiria conhecer que tipo de transporte foi usado no envio dos objectos postais (via aérea, marítima ou terrestre), matrícula do transporte entre outros dados pertinentes.

## 6.4. Algumas questões concretas e relevantes

Após implementada a *Arquitectura para um Ambiente Analítico*, e partindo de um conjunto de questões que, de acordo com o levantamento efectuado, a organização não estaria em condições de satisfazer, apresenta-se de seguida algumas dessas questões a que a nova arquitectura deve dar resposta.

- *Conheço os meus clientes?*

Com a centralização dos dados das várias aplicações operacionais de suporte ao Track & Trace no DW (SIREP), conseguem-se obter os dados relativos ao cliente, e com a ajuda de ferramentas de BI, consegue-se conhecer melhor o cliente e, inclusivamente, conhecer qual o seu comportamento futuro (com a ajuda, por exemplo, do Data Mining).

- *As iniciativas projectadas na organização estão a trazer valor?*

Esta solução permite estudar o impacto que tem uma determinada iniciativa, numa estação ou num produto. É possível analisar a quantidade vendida por dia, por estação e por produto (de acordo com o modelo apresentado). Neste caso, a gestão pode consultar indicadores diários que revelem as consequências das medidas tomadas e, caso julgue necessário, pode corrigir alguma medida tomada. A vantagem desta solução é perceber a cada momento como está o negócio e poder tomar medidas de melhoria.

- *Como conseguir aumentar o consumo dos meus clientes?*

Analisando o comportamento dos clientes, e percebendo qual ou quais os produtos que este consome, é possível fazer uma campanha promocional junto deste por forma a aumentar o seu consumo.

Esta arquitectura permite comunicar a visão estratégica da empresa, monitorizar a performance e consubstanciar as tomadas de decisão, que podem passar pela reformulação da estratégia ou passagem da visão em acções. Permite ainda identificar o valor gerado pela empresa e as componentes que afectam o resultado e fornece informação de gestão para as tomadas de decisões operacionais e estratégicas. Esta arquitectura reflecte ainda a realidade da empresa ao nível das actividades, recursos e produtos, suportando desta forma a tomada de decisões operacionais e estratégicas.

A solução apresentada (*Arquitectura para um Ambiente Analítico*) permite aos CTT a melhoria dos seus resultados no processo de retenção de clientes, na identificação de rendibilidade de projectos, permitindo a manutenção de resultados operacionais mesmo em situação de alguma diminuição de quota, e a direcção mais eficaz e eficiente de campanhas e promoções graças a uma melhor administração no contacto com o cliente e com os produtos. Através desta solução, a gestão dos CTT irá dispor da informação necessária para monitorar resultados absolutos e medir retornos nas iniciativas comerciais em desenvolvimento, devido ao facto de ter uma única fonte de dados para análise, actualizada diariamente com dados relevantes sobre clientes, facturação, circulação postal, serviços financeiros e contratação de produtos e serviços.

Em suma, permite avaliar e monitorizar a performance através de Indicadores Chave de Performance (KPI's), agir sobre esses resultados, explicando-os através das soluções que fornecem os indicadores, realizar reporting de gestão, consolidação, orçamentação e planeamento, gerir o plano e aferir o valor gerado pela empresa.

# 7. Conclusão

Na sociedade pós-industrial as organizações operavam em meios estáveis, pequenos e protegidas pelas leis do país, tinham pouca concorrência, o que as levava a estarem praticamente centradas na gestão dos seus processos. Os Sistemas Operacionais eram desenvolvidos para satisfazer as necessidades operacionais sem ter em conta as necessidades da gestão.

Este tipo de abordagem conduziu à dispersão das aplicações e dos respectivos dados uma vez que, para satisfazer necessidades contingênciais, foram surgindo aplicações dispersas pelos vários departamentos e delegações. Este crescimento desmedido levou, em muitos casos, à duplicação de aplicações, uma vez que a coordenação centralizada dos projectos informáticos não era prática corrente. Como consequência, o desenvolvimento de sistemas de informação não era realizado tendo em conta uma avaliação global das necessidades e dos sistemas já existentes na empresa. Na maioria das vezes recorria-se a entidades internas ou externas e iniciava-se uma nova aplicação, levando a um crescimento desmesurado de aplicações pela organização, consequência da falta de uma arquitectura planeada de sistemas de informação e da ausência de preocupações com a centralização de dados. Ao inverso, os dados ficavam descentralizados e cada aplicação tinha a sua forma de os armazenar. Neste panorama, não existia qualquer sistema de dados de referência na organização nem critérios de estrutura e domínio de dados que permitissem o cruzamento de dados de várias BD operacionais.

Até há pouco tempo, a situação atrás descrita era uma realidade comum. Contudo, as alterações dos mercados, provocadas pela sociedade da informação e do conhecimento, levam a uma transformação das próprias formas de gestão e à complexificação dos ambientes onde os gestores operam (ambientes complexos). Nesta nova realidade marcada pela globalização, pela desregulamentação do mercado, pela abertura de fronteiras, pelas novas tecnologias e também pelas novas ameaças e oportunidades, torna-se necessário olhar para o exterior e perceber como será possível obter vantagens competitivas.

Numa nova gestão organizacional, que se pretende mais rigorosa e competitiva, torna-se imperativo por um lado, averiguar a existência de uma aplicação que responda às necessidades identificadas, e por outro, justificar o investimento numa nova aplicação. Para que tal aconteça, a organização deve possuir ou criar um departamento de planeamento de sistemas de informação que acompanhe as mudanças e garanta as respostas adequadas. Este departamento deve ser responsável pela identificação das necessidades da organização e, nesse sentido, desenvolver sistemas de informação que reflictam um alinhamento entre as suas estratégias e funções e os objectivos de negócio da organização. Deve ainda assegurar a criação de uma arquitectura de sistemas de informação que garanta a qualidade dos dados e das aplicações.

O ambiente complexo em que grande parte das organizações passou a viver dificulta a tomada de decisão, uma vez que, por um lado, esta tem que ser rápida e eficaz, e por outro, os gestores estão conscientes que as suas decisões poderão ter repercussões a nível global. Neste sentido, torna-se pertinente que a organização esteja munida de

instrumentos específicos que apoiem a decisão. Por seu turno, a qualidade dos dados deve ser garantida tanto no seu armazenamento como na sua disponibilização, daí a importância do Data Warehouse, dos metadados e das ferramentas de Business Intelligence. Se tal acontecer, reduzem-se substancialmente as incertezas dos gestores, uma vez que terão as melhores condições possíveis para obterem informação de apoio à decisão.

Contudo, tal cenário só é possível se se verificarem diversos factores, dos quais se destacam o envolvimento de toda a organização, o total comprometimento da gestão de topo quer na comunicação, quer no apoio ao projecto de uma arquitectura para um ambiente analítico, e a partilha do conhecimento. É ainda essencial que os colaboradores estejam a par das mudanças que estão a ocorrer na organização e assim possam contribuir para o sucesso do projecto. A falta de comunicação pode desmotivar os colaboradores, comprometendo o seu contributo e, inclusivamente, levá-los a reterem o seu conhecimento.

Outro factor que pode influenciar o sucesso do projecto prende-se com a migração dos dados das várias BD operacionais para o DW. Não basta uma boa ferramenta de ETL para o efeito, é ainda necessário garantir que os responsáveis departamentais não se oponham à mudança cultural/organizacional, que está inerente à adopção da arquitectura para um ambiente analítico, nem dificultem o acesso aos dados das suas BD operacionais. Esta questão prende-se com o facto de ainda existir nas organizações uma forte resistência à mudança onde alguns colaboradores mais conservadores tendem a opor-se e a dificultar o acesso aos "***seus dados***" alegando, por exemplo, a performance dos seus sistemas. Reforça-se, então, a necessidade de explicar a todos os intervenientes

os objectivos do projecto e salientar as vantagens da implementação de um DW e respectivo ambiente analítico onde todos passam a ter acesso a dados com qualidade.

Por tudo isto, e de forma a responder à pergunta de partida deste trabalho - *Como melhorar os dados nos sistemas de apoio à decisão em ambientes complexos?* - A resposta pode ser dada da seguinte forma: *Garantido e melhorando a qualidade dos dados com a ajuda das dimensões de qualidade propostas e desenvolvendo uma arquitectura que suporte os agentes tecnológicos envolvidos e garanta a qualidade dos dados, ou seja, a arquitectura para um ambiente analítico proposta.*

Não poderíamos terminar esta dissertação sem antes deixar algumas sugestões de trabalhos futuros na área da QD como, *o estudo no sentido de encontrar métricas para associar às dimensões de qualidade dos dados propostas* e, desta forma, contribuir para melhorar os dados. Tais métricas em muito vão ajudar no processo de ETL, uma vez que vai ser possível comparar a QD para cada dimensão relativamente às métricas encontradas.

Outra sugestão prende-se com a grande importância que têm os metadados e, neste sentido, poder-se-á *estudar a qualidade dos metadados e a forma como melhorá-los*.

Por fim, ao elaboramos esta dissertação centrando-nos na proposta de uma Arquitectura para um Ambiente Analítico capaz de melhorar a QD. Deixamos o rapto para uma proposta de outra arquitectura, que responda às exigências de QD aqui estudadas, com por exemplo a *Arquitectura Orientada a Serviços (SOA)*. Esta arquitectura tem conceitos diferentes dos usados nesta dissertação como: o XML e os Web Services, embora no fundo o objectivo seja a disponibilização de dados de e com qualidade.

## Bibliografia

ALTER, Steven (1992), *Information Systems: a Management Perspective*, Addison-Wesley.

ALVES, Manuel Lopes (1995), *A Reengenharia dos Processos de Negócio*, Textos de Gestão, Lisboa, Texto Editora.

ANGELONI, Maria T. (2002), Elementos intervenientes na tomada de decisão, *Ciência da Informação*, Brasília, Vol.32, No.1, Jan./Abr., pp. 17-22.

ANTHONY, R. N. (1965), *Planning and Control Systems: A Framework for Analysis*, Harvard University Press.

ATRE, Shaku (2003a), The top 10 Critical Challenges for Business Intelligence Success, *Computerword custom publishing*, June.

ATRE, Shaku (2003b), Business Intelligence Success is Never an Accident, *Computerword custom publishing*, September.

ATRE, Shaku (1997), Learn the risks of mart, *Computerword*, Vol.31, No.20, May, pp. 63-64.

BAEZA-YATES, R. e RIBEIRO-NETO, B. (1999), *Modern Information Retrieval*, ACM Press Series, New York, Addison Wesley.

BALLARD, Chuck e HERREMAN, Dirk (1998), Data Modeling Techniques for Data Warehousing, *International Technical Support Organization*, February, IBM.

BALLOU, D. e PAZER, H. (1987), Cost/Quality Tradeoffs for Control Procedures in Information Systems, *International Journal of management Science*, Vol.15, No.6, pp. 509-521.

BARONI, Rodrigo *et al*. (2003), "Memória Organizacional" in SILVA, Ricardo V., NEVES, Ana (Org.), *Gestão de Empresas na Era do Conhecimento*, 1ª Edição, Lisboa, Edições Sílabo, pp. 212-250.

BENYON, D. (1990), *Information and Data Modelling*, Blackwell Scientific Publications.

BERSON, A. (1997), *Data Warehouse, Data Mining & OLAP*, USA, McGraw-Hill.

BOHN, Kathy (1997), Converting Data for Warehouses, *DBMS*, June, Disponível em: http://www.dbmsmag.com/9706d15.html (Acedido em Maio de 2004).

BOOCH, G., RUMBAUGH, J. e JACOBSON, I. (1999), *The Unified Modeling Language User Guide*, Addison-Wesley.

BRACKETT, M. H. (1996), *The Data Warehouse Challenge - Taming Data Chaos*, New York, John Wiley & Sons Inc.

BRAUER, Robert (2001), Data Quality is the Cornerstone of Effective BI, *Data Management Review*, October, Disponível em: http://www.dmreview.com (Acedido em Março de 2004).

CALVANESE, D., GIACOMO, G., LENZERINI, M., NARDI, D., ROSATI, R. (1997), Source Integration in Data Warehouse, *Technical Report DWQ*, Foundations of Data Warehouse Quality.

CALVANESE, D., GIACOMO, G., LENZERINI, M., NARDI, D., ROSATI, R. (2001), Data Integration in Data Warehouse, *International Journal of Cooperative Information Systems*, Vol.10, No.3, pp. 237-271.

CARREIRA, P. e GALHARDAS, H. (2004), Efficient development of data migration transformations, Disponível em: http://web.tagus.ist.utl.pt/~helena.galhardas/pub.html (Acedido em Junho de 2004).

CHEN, P. P. (1976), The Entity-Relationship Model - Toward a Unified View of Data, *ACM Transactions on Database Systems*, Vol.1, No.1, pp. 9-36.

CODD, E. F. (1970), Relational Model for Large Shared Data Banks, *Communications of the ACM*, Vol.13, No.6, pp. 377-387.

CHUCK, B., DIRK, H. e DON, S. (1998), *Data Modeling Techniques for Data Warehousing*, Red Books, IBM.

CTT (2002), *Relatório e Contas 2002*, Lisboa, Correios de Portugal.

DAELLENBACH, Hang G. (1995), *Systems and Decision Making – A management Science Approach*, John Wiley & Sons.

DATE, C. J. (1995), *An Introduction to Database Systems*, 6th edition, Addison-Wesley.

DAVENPORT, T. e PRUSAK, L. (1998), *Working Knowledge – How Organizations Manage What They Know*, Boston, MA, Harvard Business School Press.

DAVISON, Leigh (2001), Measuring Competitive Intelligence Effectiveness: Insights from the Advertising industry, *Competitive Intelligence Review*, Vol.12, No.4, pp. 25-28.

DEMIMG, W. E. (1986), *Out of the Crisis*, Cambridge University Press.

DEVLIN, Barry (1997), *Data Warehouse: from architecture to implementation*, Addison Wesley.

DONALD, B. (1997), *High Performance Oracle Data Warehousing*, USA, The Coriolis Group.

DRUCKER, P. (1996), *Sociedade pós-capitalista*, 5ª Edição, S. Paulo, Pioneira.
ECKERSON, Wayne W. (2002), Data Quality and the Bottom Line - Achieving Business Success through a Commitment to High Quality Data, *TDWI Report Series*, Disponível em: www.dw-institute.com (Acedido em Janeiro de 2004).

ENGLISH, Larry P. (1999), *Improving Data Warehouse and Business Information Quality - Methods for Reducing Costs and Increasing Profits*, New York, Wiley.

FAYYAD, U. *et al.* (1996), The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communication of ACM*, Vol.39, No.11, November, pp. 40-44.

GALHARDAS, H., FLORESCU, D., SHASHA, D. e SIMON, E. (2000), An Extensible Framework for Data Cleaning, Disponível em: http://cosmos.inesc.pt/hig/poster.pdf (Acedido em Janeiro de 2004).

GANHÃO, Fernando N. (1994), *Gestão da Qualidade*, Colecção O Gestor, IAPMEI.

GANHÃO, Fernando N. (1991), *A Qualidade Total*, Lisboa, CEDINTEC.

GARTNER GROUP (2004), Data Quality Dimentions, Disponível em http://www.gartnergroup.com (Acedido em Agosto de 2004)

GEBHARDT, M., JARKE, M., JEUSFELD, M., QUIX, C. e SKLORZ, S. (1998), Tools for Data Warehouse Quality in: *IEEE proc. Of the 10th International Conference on Scientific and Statistical Database Management*, July.

GOETSCH, D. L. e DAVIS, S. B. (1997), *Introduction to Total Quality – Quality Management for Production, Processing and Services*, Prentice Hall, pp. 1-30.

GOGGIN, John (2003), Public-Sector Ascension of the Information Maturity Model: Part1 - Data Quality, February, Meta Group, Disponivel em: http://metagroup.com (acedido em Agosto de 2003).

HACKNEY, Douglas (1998), Who Are You?, *Data Management Review*, February
Disponível em http://www.entergrupltd.com/colums/2-98-1_Who_Are_You.htm
(Acedido em Novembro de 2003).

HALL, Curt (1999), Data Warehousing for Business Intelligence, *Report for IT & Software Professionals*, March, Disponível em http://www.cutter.com/itgroup/reports/dwissues.html (Acedido em Novembro de 2003).

HERRING, Jan P. (1999), Key Intelligence Topics: A Process to Identify and Define Intelligence Needs, *Competitive Intelligence Review*, Vol.10, No.2, pp. 4-14.

HUANG, Kuan-Tsae, LEE, Yang Y. e WANG, Richard Y. (1999), *Quality Information and Knowledge*, New Jersey, Prentice Hall.

HUH, Y. U. *et al.* (1990), Data Quality, *Information and Software Technology*, Vol.32, No.8, pp. 559-565.
ILHARCO, Fernando (2003), *Filosofia da Informação – Uma introdução à informação como fundação da acção, da comunicação e da decisão*, Lisboa, Editora CampusdoSaber, Universidade Católica.

INMON, W. H. (1997), *Como construir o Data Warehouse*, 2ª Edição, Rio de Janeiro, Editora Campus.

INMON, W. H. e HACKATHORN, R. D. (1994), *Using the Data Warehouse*, New York, John Wiley & sons.

JARKE, M., JEUSFELD, M., QUIX, C. e VASSILIADIS, P. (1999), Architecture and quality in data warehouses: An extended repository approach, *Information Systems*, Vol.24, No.3, pp. 229-253.

JENNINGS, Mike (2003), The Generic Meta Data Repository, *Olap Report*, Disponível em: http://www.olapreport.com (Acedido em Fevereiro de 2004).

JEUSFELD, M., QUIX, C. e JARKE, M. (1998), *Design and Analysis of Quality nformation for Data Warehouses*, Proceedings of the 17th International Conference on Conceptual Modeling, Singapore, November, Disponível em: http://citeseer.ist.psu.edu/cache/papers/cs/12436/http:zSzzSzwww.dblab.ntua.grzSz~dwqzSzp48.pdf/jeusfeld98design.pdf (Acedido em Janeiro de 2004).

JUNQUEIRO, Raul (2002), *A Idade do Conhecimento – A Nova Era do Conhecimento*, 2ª Edição, Lisboa, Notícias Editorial.

KIMBALL, R. (1998), *The Data Warehouse Lifecycle Toolkit*, New York, John Wiley & Sons Inc.

KIMBALL, R. (1996), *The Data Warehouse Toolkit – Practical Techniques for Building Dimensional Data Warehouses,* USA, John Wiley & Sons, Inc.

KIMBALL, R. e CASERTA, J. (2004), *The Data Warehouse ETL Toolkit - Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, USA, Wiley

KIMBALL, R. e ROSS, M. (2002), *The Data Warehouse Toolkit – Guia completo para modelagem dimensional*, Brasil, *Campus.*

KOVAC, Rita, LEE, Yang W. e PIPINO, Leo L. (1997), Total Data Quality Management: The Case of IRI, Disponível em: https://www.crg2.com/iqconference/documents/publications/TDQMpub/IRITDQMCaseOct97.pdf (Acedido em Agosto 2003).

LAUDON, K. C. e LAUDON, J. P. (1998), *Management Information Systems. New Approaches to Organization & Technology*, New Jersey, Prentice Hall.

LEBARON, M. e ADELMAN, S. (1997), Meta Data Standards, *Data Management Review*, December, Disponível em: http://www.dmreview.com (Acedido em Março de 2004).

MADSEN, Mark (2004), Deciding to Buy or Build ETL for Your Data Warehouse, *Data Management Review*, October, Disponível em: http://www.dmreview.com (Acedido em Outubro de 2004).

MARCO, David (1998), Managing Meta Data, *Data Management Review*, Mars, Disponível em: http://www.dmreview.com (Acedido em Novembro de 2003).

MOHANTY, Soumendra (2004), Data Migration Strategies, part 1, *Data Management Review*, May, Disponível em: http://www.dmreview.com (Acedido em Setembro de 2004).

MOREY, Richard (1982), Estimating and improving the quality of information in a MIS, *Communications of the ACM*, Vol.25, No.6, Mai, pp. 337-342.

MOSS, Larissa T. (1998), Data Cleansing: A Dichotomy of Data Warehousing?, *Data Management Review*, February, Disponível em: http://www.dmreview.com (Acedido em Junho de 2004).

MOSS, Larissa T. e ATRE Shaku (2003), *Business Intelligence Roadmap – The Complete Project Lifecycle for Decision-Support Application*, USA, Addison-Wesley.

NOLAN, Richard (1995), *Creative Destruction: A Six Stage for Transforming the Organization*, Harvard Business School Press.

NONAKA, I. e TAKEUCHI, H. (1995), The Knowledge Creating Company, New York, Oxford University Press.

ORACLE (2004), Oracle Notícias, Disponível em: http://www.oracle.pt (Acedido em Setembro 2004).

ORR, Ken (1998), Data Quality and Systems Theory, *Communications of the ACM*, Vol.41, No.2, February, pp. 66-71.

PALMA-DOS-REIS, A. (1999), *Sistemas de Decisão*, Lisboa, Universidade Aberta.

PEREIRA, José Luís (1998), *Tecnologias de Base de Dados*, Tecnologias de Informação, 3ª Edição, FCA.

PIERCE, Elizabeth M. (2004), Assessing Data Quality UIT Control Matrices, *Communications of the ACM*, Vol.47, No.2, February, pp.82-86.

PIPINO, Leo L., LEE, Yang W. e WANG, Richard Y. (2002), Data Quality Assessment, *Communications of the ACM*, Vol.45, No.4, April, pp.211-218.

PIRES, A. R. (2000), *Qualidade – Sistema de Gestão da Qualidade*, 2ª Edição, Lisboa, Edições Sílabo.

POE, Vidette *et al*. (1998), *Building a Data Warehouse for Decision Support*, 2ª Edition, New Jersey, Prentice Hall.

QUINN, James B. (1992), *Intelligent entreprise*, FreePress.

RAHM, E. e DO H. (2000), Data Cleaning: Problems and Current Approaches, *Bulletin of the Technical Committee on Data Engineering*, Vol.23, No.4, Disponível em: http://research.microsoft.com (Acedido em Julho de 2004).

RASCÃO, José (2001), *Análise Estratégica - Sistemas de Informação para a Tomada de Decisão Estratégica*, 1ª Edição, Lisboa, Edições Sílabo.

RASCÃO, José (2000), *Sistemas de Informação para as Organizações – A Informação Chave para a Tomada de Decisão*, 1ª Edição, Lisboa, Edições Sílabo.

REDMAN, Thomas (2004), Data: An Unfolding Quality Disaster, *Data Management Review*, August, Disponível em: http://www.dmreview.com (Acedido em Setembro de 2004).

REDMAN, Thomas (1998), The impact of poor quality data on the typical enterprise, *Communications of the ACM*, Vol.41, No.2, February, pp. 79-82.

REDMAN, Thomas (1996), *Data Quality for the Information Age*, Boston, MA, Artech House.

REZENDE, Yara (2002), Informação para negócios: os novos agentes do conhecimento e a gestão do capital intelectual, *Ciência da Informação*, Brasília, Vol.31, No.1, Jan./Abr., pp. 75-83.

RODRIGUES, Luís S. (2002), *Arquitecturas de Sistemas de Informação*, Lisboa, FCA.

RUBIN, Jon (2003), The DB2 Framework for Business Intelligence, *Information Management Software*, May, IBM, Disponível em: http://www.ibm.com/software/data/integration (acedido em Julho 2004).

SAS Institute (2004), Intelligence Value Chain: Business Intelligence, *SAS White Paper*, Disponível em: http://www.sas.com.pt (acedido em Junho 2004).

SATYA, Sachdeva (1998), Meta Data Architecture for Data Warehousing, *Data Management Review*, April, Disponível em: http://www.dmreview.com (Acedido em Março de 2004).

SENGE, P. (1990), *The Fifth Discipline – The art & Practice of Learning Organization*, London, Century Business.

SELIGMAN, L. e ROSENTHAL, A. (1996), *A Metadata Resource to Promote Data Integration*, IEEE Metadata Conference, Silver Spring, April, Disponível em: http://www.mitre.org/tech/itc/staffpages/arnie/pubs/ (Acedido em Setembro de 2003).
SHERMAN, R. (1997), Metadata: The Missing Link, *DBMS*, August.

SIMON, H. A. (1960), *The new science of management decision*, New York, Harper & Row.

SINGH, Harry (1997), *Data Warehousing: Concepts, Technologies, Implementation, and Management*, Upper Saddle River, Prentice Hall.

SILVA, Ricardo V., NEVES, Ana (Org.) (2003), *Gestão de Empresas na Era do Conhecimento*, 1ª Edição, Lisboa, Edições Sílabo.

SOUSA-MENDES, Aristides (2001a), *A Qualidade dos Dados nos Sistemas de Informação*, Lisboa, Dissertação de mestrado apresentada à Universidade Católica Portuguesa.

SOUSA-MENDES, Aristides (2001b), Sociedade da Informação ou Sociedade do Conhecimento?, *Revista Portuguesa de Gestão*, Lisboa, Novembro, pp. 16-25.

STROG, Diana, LEE, Yang Y. e WANG, Richard Y. (1997), Data Quality in Context, *Communications of the ACM*, Vol.40, No.5, May, pp.103-110.

TABORDA, João P. e FERREIRA, Miguel D. (2002), *Competitive Intelligence – Conceitos, Práticas e Benefícios*, Empresa Inteligente, 1ª Edição, Lisboa, Pergaminho.

TIWANA, A. (2000), *The knowledge management toolkit*, Prentice-Hall.

TURBAN, Efrain (1995), *Decision Support Systems and Expert Systems –Management Support Systems*, 4ª Edition, New Jersey, Prentice-Hall.

TURBAN, Efrain e ARONON, Jay (1998), *Decision Support Systems and Intelligent Systems*, 5ª Edition, New Jersey, Prentice-Hall.

TURBAN, Efrain e MEREDITH, Jack (1994), *Fundamentals of Management Science*, 6ª Edition, IRWIN.

VARAJÃO, João E. Q. (1988), *A Arquitectura da Gestão de Sistemas de Informação*, 2ª Edição, Lisboa, FCA.

VASSILIADIS, P., BOUZEGHOUB, M. e QUIX, C. (1999), Towards Quality-Oriented Data Warehouse Usage and Evolution, Disponível em: http://www.dbnet.ece.ntua.gr/~dwq/p41.pdf (Acedido em Setembro de 2004).

VASSILIADIS, P., SIMITSIS, A. e SKIADOPOULOS, S., (2002), Conceptual Modeling for ETL Processes, Disponível em: http://citeseer.ist.psu.edu/cache/papers/cs/28930/http:zSzzSzwww.dblab.ece.ntua.grzSz~pvassilzSz.zSzpublicationszSzdolap02_CR.pdf/vassiliadis02conceptual.pdf (Acedido em Setembro de 2004).

VELHO, Amândio Vaz (2004), *Arquitectura de Empresa*, 1ª Edição, Lisboa, Centro Atlântico.

VITT, Elizabeth *et al.* (2002), *Business Intelligence: Making Better Decisions Faster*, Microsoft Press.

WALZER, Peter (2004), Why So Many Business Intelligence Initiatives Fail, *Data Management Review*, February, Disponível em: http://www.dmreview.com (Acedido em Setembro de 2004).

WAND, Yair e WANG, Richard Y. (1996), Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, Vol.39, No.11, November, pp. 86-95.

WANG, Richard Y. (1998), A Product Perspective on Total Data Quality Management, *Communications of the ACM*, Vol.41, No.2, February, pp. 58-65.

WANG, Richard Y. *et al.* (1995), Toward quality data: An attribute-based approach, *Decision Support Systems*, No.13, pp. 349-372.

WANG, R. Y., STOREY, V. C. e FIRTH, C. P. (1995), A framework for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering*, Vol.7, No.4, August, pp. 623-640.

WANG, Richard Y. e STRONG, Diana (1996), Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, No.4, pp. 5-34.

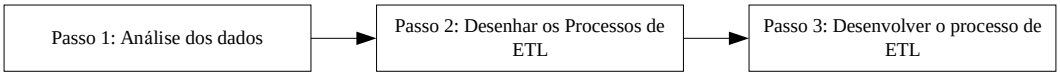
WATSON, Hugh J. (1998), Managerial Considerations, *Communications of the ACM*, Vol.41, No.9, September, pp. 32-37.

WEBBER, A. (1993), What's so New About the New Economy, *Harvard Business Review*, Jan./Feb., pp. 22-42.

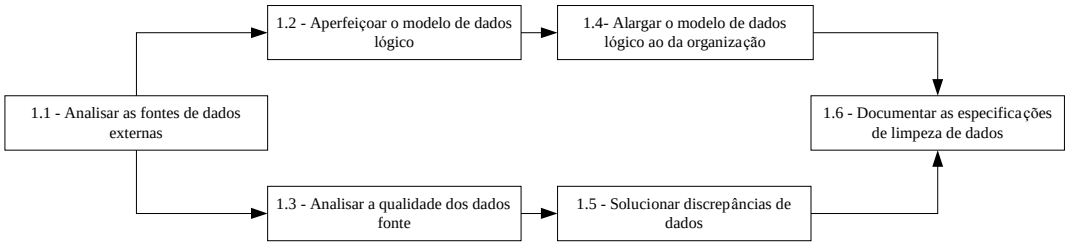
WINTER, Richard (1999), Be Aggregate Aware, *Intelligent Enterprise Magazine*, Vol.2, No.13, September, Disponível em: http://www.intelligententerprise.com/db_area/archives/1999/991409/scalable.jhtml (Acedido em Novembro de 2003).

WHITE, Colin J. (2003), Corporate Performance Optimization Guide, *Intelligence Business Strategies*, Version 2, January, Sponsored by Oracle Corporation, Disponível em: http://www.intelligentbusiness.biz (Acedido em Março de 2004).

# Anexo 1 - Passos para a Análise dos dados nos sistemas fonte e, Desenho e Desenvolvimento dos processos de ETL




## Passo 1: Análise dos dados




**1.1 - Analisar as fontes externas de dados**

Identificar as entidades e relações de cada fonte de dados externa
Fundir as novas entidades e relações das fontes de dados externas no modelo de dados lógico

**1.2 - Aperfeiçoar o modelo de dados lógico**

Criar todos os atributos do modelo lógico para incluir os elementos necessários quer de fontes de dados internos, quer externos
Criar novas entidades e relações onde seja necessário, para armazenar os novos atributos
Analisar o layout das fontes identificadas
Analisar o conteúdo das fontes identificadas

**1.3 - Analisar a qualidade dos dados fonte**

Aplicar regras de integridade e de domínio, para validar dados inválidos, tais como:
*valores em falta, valores perdidos, valores escondidos, valores contraditórios, valores que violam as regras de negócio, falta de chaves primárias, chaves primárias duplicadas*
Determinar a severidade do problema (quantificar as anomalias identificadas)
Determinar a critícidade do problema (qual o impacto das anomalias)

**1.4 - Alargar o modelo de dados lógico ao da organização**

Fundir o modelo de dados lógicos do projecto no modelo de dados lógico da organização
Identificar discrepâncias e inconsistências de dados entre os modelos de dados lógicos

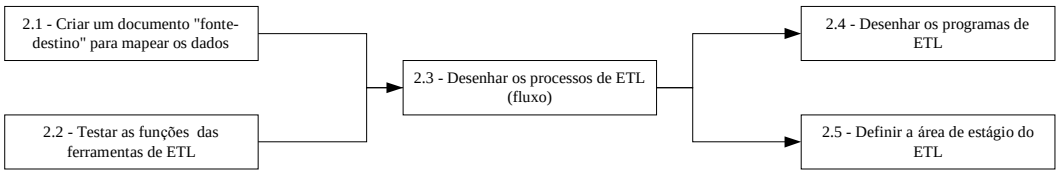
**1.5 - Solucionar discrepâncias de dados**

Discutir as discrepâncias dos dados com os responsáveis
Ajustar o modelo de dados lógicos do projecto ao modelo de dados da organização
Notificar outras equipas de projecto que possam ser afectadas
Documentar as discrepâncias dos dados

**1.6 - Documentar as especificações de limpeza de dados**

Rever a classificação de elementos de dados, como: crítico, importante, insignificante
Documentar as especificações de limpeza dos dados para os dados críticos
Documentar as especificações de limpeza dos dados para os outros dados seleccionados

## Passo 2: Desenhar os Processos de ETL




**2.1 - Criar um documento "fonte-destino" para mapear os dados**

Rever os planos para cada fonte de dados externos
Rever a descrição de dados cada fonte de dados externos
Rever as especificações de limpeza de dados para as fontes de dados externos
Criar uma matriz para todas as tabelas alvo
Listar todas as fontes de dados a usar (Base de Dados ou ficheiros)
Listar todas as fontes de dados a usar (Tabelas)
Documentar as especificações de transformação:
Conjugar o conteúdo dos dados das múltiplas fontes (se necessário)
Dividir o conteúdo dos dados por colunas (se necessário)
Incluir algoritmos de agregação e de sumarização
Incluir especificações de limpeza de dados para cada coluna
Considerar a integridade referencial (se não estiver contemplada pelo DBMS)
Considerar mensagens de erro, como a rejeição de registos
Considerar a reconciliação (número de registos, domínio)

**2.2 - Testar as funções das ferramentas de ETL**

Rever as especificações de transformação no documento "fonte-destino" dos dados
Determinar se as funções das ferramentas de ETL podem executar a lógica de transformação exigida
Determinar se é necessário escrever código adicional

**2.3 - Desenhar os processos de ETL (fluxo)**

Determinar a sequência mais eficiente para extrair os dados das fontes
Determinar a sequência mais eficiente para transformar, limpar, e carregar os dados
Identificar todos os ficheiros de trabalho temporários e todas as tabelas
Determinar que componentes do processo de ETL podem correr em paralelo
Determinar que tabelas podem ser carregadas em paralelo
Desenhe o diagrama de fluxo de processos:
*extracção de dados dos sistemas fonte, tabelas e ficheiros temporários e permanentes*
*programas de transformação, registo de erros e geração de relatórios de erro,*
*carregamento de dados*

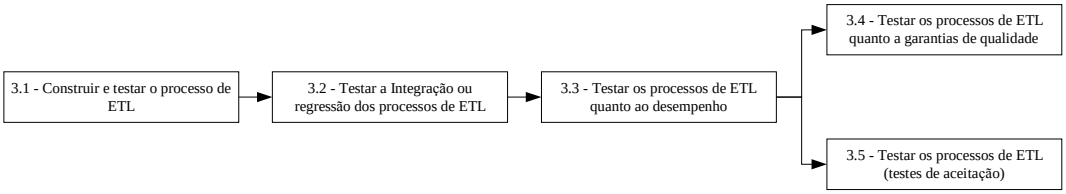
**2.4 - Desenhar os programas de ETL**

Optimizar os processos de ETL
Traduzir as especificações de transformação em programas específicos

**2.5 - Definir a área de estágio do ETL**

Determinar como distribuir o processo de ETL
Reservar espaço para ficheiros e tabelas de trabalho temporários e permanentes
Criar bibliotecas (rotinas/procedimentos)
Estabelecer um controlo de versões para os procedimentos

## Passo 3: Desenvolver o processo de ETL




**3.1 - Construir e testar o processo de ETL**

Codificar os programas de ETL
Se usar uma ferramenta de ETL, escrever instruções para os diferentes módulos
Capturar os metadados gerados para o repositório de metadados
Escrever código no programa de ETL de forma a produzir totais, métrica de qualidade e estatísticas de carga
Testar individualmente cada módulo do programa
Se usar uma ferramenta de ETL, testar cada módulo da ferramenta
Escrever o código para executar o programa ETL

**3.2 - Testar a Integração ou regressão dos processos de ETL**

Criar um plano de testes para testar os processos de ETL
Obter dados para teste
Documentar os testes
Comparar os resultados dos testes actuais com os valores esperados
Rever os programas de ETL (ou os processos de ETL)
Testar novamente todo o processo de ETL (do princípio ao fim)

**3.3 - Testar os processos de ETL quanto ao desempenho**

Testar os programas ou módulos de ETL que leram ou escreveram nas tabelas de maior volume
Testar a execução paralela de programas ou módulos de ETL contra as tabelas de maior volume
Testar os programas ou módulos de ETL que executam operações complicadas
Usar grandes quantidades de dados para testar o desempenho (testes de simulação)

**3.4 - Testar os processos de ETL quanto a garantias de qualidade (Quality Assurance)**

Submeter os programas de ETL ao ambiente de garantia de qualidade
Testar o processo de ETL do princípio ao fim
Obter aprovação do pessoal das operações para colocar os processos de ETL em produção

**3.5 - Testar os processos de ETL (testes de aceitação)**

Testar o processo de ETL do princípio ao fim
  Validar todas as transformações de limpeza
  Validar as rotinas
  Validar os totais de reconciliação
Obter a certificação para o processo de ETL